Genetic epidemiology of markers of genomic ageing

Chen Li

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College

November 2019

Preface

This dissertation aims to characterize genomic ageing using genetic and observational epidemiological approaches, with an emphasis on two markers of genomic ageing, leukocyte telomere length and mosaic loss of chromosome Y, providing insights into their biological mechanisms and clinical relevance. Besides the work I described here, I have contributed to other collaboration projects outside of the scope of this dissertation during the course of my PhD, including a genetic discovery of human plasma metabolome and iron metabolism, gene-specific effects of low-density-lipoprotein cholesterol on type 2 diabetes and phenome-wide association studies across multiple traits in UK biobank. Much of this work has been written into scientific manuscripts, under review or published, which are listed in Appendix B.

I declare that the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at the University of Cambridge or any other University or similar institution. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the Acknowledgements or main text. This dissertation contains fewer than 60,000 words excluding figures, tables, appendices and references.

Chen Li

November 2019

Abstract

Name Chen Li

Title Genetic epidemiology of markers of genomic ageing

Background Ageing is associated with changes in physical functioning, generally leading to a progressive decline in health and development of age-related diseases. Age-related changes also affect our genome and markers of genomic ageing, such as telomere length and chromosomal loss, have been linked to cancer. The genetic architecture of these markers is not well understood and studies investigating associations with common age-related cardiometabolic conditions have been limited in their design, analytical methods, power and genetic instruments used. Only a few studies have investigated prospective changes in these markers with age.

Objectives To study the epidemiology of two heritable markers of genomic ageing, leukocyte telomere length (LTL) and mosaic loss of chromosomal Y (mLOY) and test their causal relevance for cardiometabolic and other age-related disorders.

Methods Large-scale, genome-wide meta-analysis, two-sample Mendelian Randomization (MR), and prospective observational case-cohort analysis methods were used to (1) identify novel genetic determinants of LTL, (2) investigate causal associations of genetic differences in LTL with disease, (3) review the evidence on and assess the feasibility of studying longitudinal changes of LTL, and (4) assess observational associations between LTL and mLOY and future risk of type 2 diabetes (T2D) in a large, international case-cohort study.

Results Genome-wide meta-analyses including 78,592 individuals identified 49 regions associated with LTL at FDR<0.05 including 17 (6 novel) at *p*-value<$5X10^{-8}$. A total of 32 candidate genes were prioritised with strong suggestive evidence for their roles in telomere homeostasis, DNA repair and nucleotide metabolism. Targeted and phenome-wide MR analyses suggested causal associations of shorter LTL with an increased risk of cardiovascular conditions, and decreased risks of multiple cancer types and diseases of excessive growth. LTL shortening was observed even in young and healthy individuals, and baseline LTL was strongly associated with the rate of shortening, questioning the usefulness of LTL shortening rate as an outcome in genetic association studies. No evidence was found for strong associations of mLOY or LTL either measured or genetically predicted with the risk of T2D.

Conclusion Our findings substantially expand current knowledge on genes and mechanisms regulating LTL, as well as refine our understanding of the impact of genetic differences in LTL on human health and disease, while providing no strong evidence for prospective observational or causal associations between markers of genomic ageing and T2D risk.

I would like to dedicate this thesis to my parents.

Acknowledgements

Firstly, I would like to thank my supervisor Dr Claudia Langenberg for her support and guidance throughout the entire course of my PhD. She has been a kind, responsible and inspirational mentor, guiding me onto the right path of becoming an independent and innovative researcher. I would also like to thank Prof Nick Wareham for the opportunity to work on a series of exciting novel projects. His vision and leadership have positive impact on constructing the work from wider scientific perspectives.

There are a number of other senior scientists in the MRC Epidemiology Unit, who provided great support to the work presented in this dissertation and their support are much appreciated. Specifically, I would like to thank Dr Luca Lotta for his insightful guidance and leadership in developing various collaboration projects listed in the Appendix B; Dr Isobel Stewart for her guidance and kind support across various projects, especially the Metabolon project; Dr Felix Day for his kind support and advice on various technical issues and career development; Dr John Perry for his contributions and comments on the project of loss of chromosome Y; and Dr Fumiaki Imamura for his guidance and help in technical questions; and Drs Ken Ong and Nita Forouhi for opportunities of collaborating on various projects across epidemiological fields.

I would also like to acknowledge the operational and laboratory team at the MRC Epidemiology Unit for sample logistics and processing and running of telomere measurements; the data management team for managing the EPIC-InterAct data; the statistics and data science team for providing technical support for statistical analyses.

During the course of my PhD, I led a genome-wide association meta-analysis study of leukocyte telomere length (Chapter 2 and 3). This was conducted mainly in collaboration with the ENGAGE telomere working group based at the University of Leicester, who helped to advance the progress of the analyses and completion of the manuscript. I would like to thank all of the members of the group, especially Dr Veryan Codd, Dr Christopher Nelson and Prof Nilesh Samani. I would also like to thank other two collaboration teams on this project, our Cambridge collaborators from the Cardiovascular Epidemiology Unit who established and managed the EPIC-CVD case-cohort study, especially Dr Adam Butterworth, Tao Jiang and Prof John Danesh; and the American team at the Scripps Research Institute, Taylor Loe and Dr Eros

Lazzerini Denchi (now at NIH), who helped to perform the cell model-based functional analyses. I performed all the analyses for this thesis, produced figures and tables, and drafted the text. Parts of chapters 2 and 3 are part of a manuscript that was done in collaboration with Engage investigators and that I am first author of (under review; Appendix B)

I would also like to thank several individuals who contributed specifically to certain chapters, the details of which are provided below.

The laboratory and data management team at the MRC Epidemiology Unit helped to conduct longitudinal measurement of leukocyte telomere length in a pilot study of around 100 samples. Particularly, I would like to thank Dr Debora Lucarelli, Lucy Finnegan, Vasileios Kaimakis and Nicola Kerrison for their contributions to performing experimental analyses and quality control, experimental protocol optimisation and raw data processing in the Chapter 4. I analysed the data and drafted the text. All tables and figures were produced by me, except those for summarising experimental approaches and results, which were adapted from a laboratory report generated by the laboratory and data management team.

The analysis plan for the work presented in Chapter 5 was developed in consultation with Stephen Sharp who provided helpful comments on statistical techniques. I analysed the data myself, produced all tables and figures, and drafted the text.

Lastly, I would like to thank my family for immense encouragement and unwavering support throughout the entire journey and beyond, especially my partner, Dr Ben Sun for reading through the thesis and providing helpful feedback and suggestions.

TABLE OF CONTENTS

List of Figures

List of Tables

List of supplementary information

Supplementary notes, tables and figures are shown in the Appendix A.

Supplementary Notes

Information on Study Cohorts of LTL GWAS meta-analysis
The demographic characteristics of all study cohorts, for both discovery and replication phases are shown in the Supplementary Table 1. All individuals included in the analysis are of European descent.

Description of Individual loci associated with LTL
Detailed description of FDR loci identified in the LTL GWAS meta-analysis. Evidence for prioritisation of likely causal genes, including functionality of SNPs, bioinformatic support and literature review. References were listed in the Supplementary References.

Systematic literature review on longitudinal changes of LTL
Searching strategies applied and summary of the study results.

Supplementary Figures

Supplementary Figure 1: Study design. Schematic graph to illustrate study design of the LTL GWAS meta-analysis. GWAS was conducted in each individual study cohort, stratified by genotyping platform and disease status. SNP genotyping, GWAS and meta-analyses as well as the corresponding QC procedures were described in detail in sections 2.2.3 and 2.2.4.

Supplementary Figure 2: Manhattan Plot. Manhattan plot with quantile-quantile plot inlay. Known loci were labelled in blue, novel loci associated with LTL at genome-wide significance (p-value<$5x10^{-8}$, red line) in red, and at FDR threshold of 5% (blue line) in orange.

Supplementary Figure 3: Regional plots of genome-wide significant loci (regions around conditionally independent lead variants). Regional plots illustrate 400kb windows encompassing conditionally independent variants, except the *TERT* locus which is illustrated as a 200kb window.

Supplementary Figure 4: Distributions of mLRRY values in **A.** EPIC-InterAct and **B**. UK biobank, before (left) and after (right) data transformation. Z_invn_mL_n means standardised values of mLRRY after a series of data transformation (winsorisation at 5SD, followed by inverse normal transformation and z-standardisation).

Supplementary Figure 5: Distribution of mLRRY values in each EPIC-InterAct participating country separately, before (upper) and after (bottom) data transformation.

Supplementary Figure 6: Observational associations between mLRRY and T2D risk. Same models were applied as described in the Figure 5.1, with association estimates shown in each

country. *Mdiet*: Mediterranean diet score, *alc*: lifetime alcohol consumption, *pa*: physical activity, ed: educational level, *bmi*: body mass index, *wc*: waist circumference.

Supplementary Figure 7: Observational associations between mLRRY and T2D risk. Same models were applied as described in the Figure 5.1, but with mLRRY as a binary variable (mLRRY<0, coded as 0, i.e. indicating mLOY).

Supplementary Tables

Supplementary Table 1: Cohort demographics and LTL measurement data. T/S distributions are given from primary data prior to z-transformation. Level of statistical significance is denoted by *p<0.01, **p<0.0001. All cohorts showed expected age-associated decline in LTL and higher LTL in women compared to men, except In FINNRISK and NTR_GO2 cohorts, the gender effect was not significant, most likely due to small sample sizes. For the measurement laboratory: 1, Leicester; 2, Helsinki; 3, London; 4, Genetic Laboratory Erasmus MC, Rotterdam; 5, laboratory of Telomere Diagnostics Inc., CA, USA; 6, Cambridge. The inter-run coefficient of variation (CV) is given for LTL measurements performed on triplicates of the same samples.

Supplementary Table 2: Details of genotyping platforms and analysis methods used by each study.

Supplementary Table 3: LD between sentinel variants for previously reported loci. LD (R^2 and D') were calculated using LDLink (https://ldlink.nci.nih.gov) between sentinel variants identified in this study and those previously reported. These are broken down by ancestry of the populations from reported studies. LD is calculated for both Europeans (CEU) and for the reported ancestries (CHS or BEB) based on 1000 genomes information.

Supplementary Table 4: Independent variants associated with LTL at FDR<0.05. Columns indicate (Chr) chromosome ; SNP; (bp) physical position (hg19); (freq) frequency of the effect allele in the original GWAS data; (refA) the effect allele; (b) effect size, (se) standard error and (p) *p*-value from single variant based GWAS meta-analysis; (n) estimated effective sample size; (freq_geno) frequency of the effect allele in the reference sample; (bJ),(bJ_se), (pJ) effect size, standard error and *p*-value from joint models; and (LD_r) between the variant and the locus sentinel variant.

Supplementary Table 5: Comparison of all loci at FDR<0.05 to that reported in the Singaporean Chinese Health Study (SCHS). Data is sorted by original *p*-value, pJ indicates *p*-value from conditional (GCTA) analyses. Minor allele frequencies (MAF) are given from 1000 genomes populations for information. Variants with MAF<0.01 were excluded in the SCHS study so not available. Many of our variants were monoallelic in the SCHS and denoted by" - ". Variants that were only genotyped in our study but not in the SCHS dataset or 1000 genomes reference panel, were denoted by "NA".

Supplementary Table 6: Functional prediction of nonsynonymous variants. Coding variants were identified within each locus with $r^2 \geqslant 0.8$ to the locus lead SNP. Functional prediction of the amino acid changes was carried out using PolyPhen, SIFT and CADD prediction tools. CADD

scores above 20 are considered to be within the 1% most deleterious mutations. PD: probably damaging; B: benign; U: unknown; T: tolerant; D: damaging.

Supplementary Table 7: Integration of eQTLs using S-PrediXcan and co-localisation analyses. Genes are identified by Ensembl IDs and gene names are derived from the UCSC Human Genome database. Genes were allocated to overlapping LTL loci where possible, with sentinel SNPs of the corresponding loci shown. Detailed column specifications were given in software websites (section 2.2.6.2).

Supplementary Table 8: Integrated scoring of non-coding variants. Scoring was performed with SNP Nexus IW scoring tool.

Supplementary Table 9: Identification of meQTLs. Independent SNPs associated with LTL at FDR<0.05 and their proxies (r^2<0.8) were searched in meQTL databases using PhenoScanner (section 2.2.6.3). Best proxy SNPs were those that exhibited the highest LD r^2 with locus sentinel SNPs; the corresponding rows indicate their associations with DNA methylation markers. Most significant meQTLs indicate SNPs that were most significantly associated with DNA methylation markers within each independent LTL signal, and their blocks show their associations with the DNA methylation markers and LD r^2 with the independent LTL signal SNPs.

Supplementary Table 10: Gene prioritisation. Evidence to support likely-causal genes, including nonsynonymous variants, eQTLs, known roles in telomere regulation and having other supportive information from literature. Genes were prioritised based on most lines of evidence or on strength of evidence (including deleteriously predicted mutations, known roles in telomere biology and eQTLs in multiple tissues over a single tissue).

Supplementary Table 11: Pathway analysis. Prioritized genes or the closest genes to locus sentinel variants where no prioritization was possible were used as input to PANTHER (section 2.2.8.1). A statistical over-representation analysis was performed. Pathways over-represented at FDR<0.05 are shown.

Supplementary Table 12: LD score regression (*p*-value<0.05). Genome-wide genetic correlations between LTL and different traits.

Supplementary Table 13: Case definition for 122 diseases manually curated within UK Biobank.

Supplementary Table 14: Estimated power to detect an odds ratio (OR) in the range of 0.9 to 1.1 for given numbers of cases within UK Biobank.

Supplementary Table 15: Significant associations (*p*-value < 0.05) between genetically predicted LTL and diseases among 122 diseases manually curated in UK Biobank. Nominally significant associations were highlighted in yellow, among which that passed the Bonferroni corrected significance threshold were in red.

Supplementary Table 16: Definitions of 27 cancers based on self-reported disease histories and ICD-10 codes in UK Biobank.

Supplementary Table 17: Diseases associated with genetically predicted LTL at Bonferroni-corrected significance (*p*-value (IVW) < $1.3X10^{-4}$).

Supplementary Table 18: Distributions of the mLRRY values in each EPIC-InterAct country, separately, and overall. Distributions before (upper) and after (bottom) data transformation are shown.

Supplementary Table 19: Observational associations between mLOY (binary, mLRRY<0) and T2D risk in UK Biobank. Associations were analysed using logistic or Cox regression models for prevalent and incident T2D cases, respectively, with different adjustments, as shown in the table.

Supplementary Table 20: Age or smoking stratification analyses in UK Biobank. Associations were analysed using logistic or Cox regression models for prevalent and incident T2D cases, respectively. Models were adjusted for centre and array in the age-band stratified analyses, and additionally for age in the smoking stratified analyses.

Supplementary Table 21: Meta-regression analyses to identify sources of heterogeneity for associations between mLRRY and T2D risk. Smoking status and age band were analysed separately, i.e. individuals were stratified by country and smoking status (ever vs. never) or by country and age band (<50, 50-65 and >65), resulting in 14 and 18 strata, respectively. In each stratified analysis, beta coefficients were combined across strata using random-effects meta-regression models. variances between strata (tau^2) were estimated by the residual (restricted) maximum likelihood (REML) algorithm with Knapp and Hartung modification to control type I error. In addition, permutation-based *p*-values were calculated, either with or without adjustment for multiple testing.

Supplementary Table 22: Missingness of mLRRY in EPIC-InterAct study. **A**. Proportion of mLRRY missingness in each country. **B**. Proportions of mLRRY missingness in T2D incident cases and controls. **C**. Age distributions among individuals with or without mLOY measurements. **D**. Factors associated with the missingness of mLRRY.

List of abbreviations

A	Adenine
ACYP2	Acylphosphatase 2
AD	Alzheimer's Disease
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
AMPK	Adenosine Monophosphate-activated Protein Kinase
APOE	Apolipoprotein E
ATP	Adenosine triphosphate
AUC	Area Under the Curve
ROC	Receiver-Operator Curve
BAF	B Allele Frequency
BAG6	BCL2 Associated Athanogene 6
BMI	Body mass index
BRCA	Breast Cancer Susceptibility Protein
C	Cytosine
c-NHEJ	classical Non-Homologous End Joining
CAD	Coronary Artery Disease
CARDIoGRAM	Coronary ARtery DIsease Genome wide Replication and Meta-analysis
CARMIL1	Capping Protein Regulator And Myosin 1 Linker 1
CSNK2B	Casein Kinase 2 Beta
CCH	Copenhagen City Heart
CHD	Coronary Heart Disease
CHRN	Cholinergic Receptor Nicotinic
CNV	Copy Number Variation
COPD	Chronic Obstructive Pulmonary Disease
CTC1	CST Telomere Replication Complex Component 1
CTCF	CCCTC-binding Factor
CVD	Cardiovascular Disease
dA	deoxyadenosine
dAMP	deoxyadenosine monophosphate
dC	deoxycytidine
DCAF4	DDB1 And CUL4 Associated Factor 4
DCK	Deoxycytidine Kinase
dCMP	deoxycytidine monophosphate
DDR	DNA Damage Response
dG	deoxyguanosine
dGMP	deoxyguanosine monophosphate
DHX35	DEAH-Box Helicase 35
DKC1	Dyskerin Pseudouridine Synthase 1
DNA	Deoxyribonucleic Acid

dNTP	deoxyribonucleoside triphosphate
dT	deoxythymidine
dTMP	deoxythymidine monophosphate
ENGAGE	European Network for Genetic and Genomic Epidemiology
EPIC	European Prospective Investigation of Cancer
FDR	False Discovery Rate
FOXO	Forkhead box O
G	Guanine
GCTA	Genome-wide Complex Trait Analysis
GH	Growth Hormone
GWA	Genome-Wide Association
GWAS	Genome-Wide Association Study
HLA	Human Leukocyte Antigen
HRC	Haplotype Reference Consortium
HR	Hazard Ratio
ICD9/10	the 9th/10th revision of the WHO International Classification of Diseases
IGF	Insulin/Insulin-like Growth Factor
IL6	Interleukin 6
LD	Linkage Disequilibrium
LDSC	LD score regression
LMNA	Lamin A/C
LPA	Lipoprotein(a)
LRR	$\log_2(R_{observed}/R_{expected})$, R: signal intensity
LRRC16A	Leucine-Rich Repeat-Containing Protein 16A
LTL	Leukocyte Telomere Length
MAF	Minor Allele Frequency
MHC	Major Histocompatibility Complex
mLOX	mosaic Loss Of chromosome X
mLOY	mosaic Loss Of chromosome Y
mLRRY	median of LRR of chromosome Y
MOB1B	MOB Kinase Activator 1B
MPHOSPH6	M-Phase Phosphoprotein 6
MR	Mendelian Randomisation
mTOR	mechanistic Target Of Rapamycin
mTORC1	mTOR Complex 1
NAD	Nicotinamide Adenine Dinucleotide
NAF1	Nuclear Assembly Factor 1
OBFC1	Oligonucleotide/Oligosaccharide-Binding Fold-Containing Protein 1
OR	Odds Ratio
PARP1	Poly(ADP-Ribose) Polymerase 1
PCR	Polymerase Chain Reaction
PC	Principal Component

PheWAS	Phenome-wide Association Study
POT1	Protection of Telomeres Protein 1
PP	Posterior Probability
PPA	Posterior Probability of Association
PREVEND	Prevention of REnal and Vascular ENd stage Disease
PRRC2A	Proline Rich Coiled-Coil 2A
PRS	Polygenic Risk Score
PXK	PX Domain Containing Serine/Threonine Kinase Like
QC	Quality Control
qPCR	quantitative PCR
QQ	Quantile–Quantile
QTL	Quantitative Trait Loci
RAP1	Ras-related Protein Rap-1
RCT	Randomised Controlled Trail
RFWD3	Ring Finger And WD Repeat Domain 3
RNR	Ribonucleotide Reductase
RTEL1	Regulator of Telomere Elongation Helicase 1
SAMHD1	SAM and HD Domain Containing Deoxynucleoside Triphosphate Triphosphohydrolase 1
SD	Standard Deviation
SENP7	SUMO Specific Peptidase 7
SMUG1	Single-strand Selective Monofunctional Uracil DNA Glycosylase
SNP	Single Nucleotide Polymorphism
T	Thymine
T2D	Type 2 Diabetes
TEN1	TEN1 Subunit Of CST Complex
TERC	Telomerase RNA Component
TERF	Telomeric Repeat Binding Factor
TERT	Telomerase Reverse Transcriptase
TFBS	Transcription Factor Binding Sites
TINF2	TERF1 Interacting Nuclear Factor 2
TK1	Thymidine Kinase
TL	Telomere Length
TPP1	Tripeptidyl Peptidase 1
TYMS	Thymidylate Synthetase
VEP	Variant Effect Predictor
ZNF	Zinc Finger Protein

Chapter 1

Introduction and literature review

Ageing is regulated by a dynamic interplay between genes and environment, which influences a number of cellular hallmarks, including genome integrity and telomere length (TL) homeostasis[1,2]. Genetic mutations of key genes governing these processes have been linked to age-related diseases and lifespan. However, understanding of genetic contribution to these hallmarks of ageing has so far largely been driven by experimental studies in short-lived non-vertebrate model organisms, whereas in complex vertebrate systems, notably in humans, the genetic regulation of these ageing hallmarks is under-studied. Studying human genetic variation underlying dysregulation of these hallmarks of ageing may facilitate discovery of genes and signalling pathways that underpin natural ageing processes. Investigation of associations of these genetic variants with clinical outcomes can facilitate protection from disease development, thereby contributing to healthy longevity. In this chapter, I will first introduce background and methodology for studying human genome variation, with a literature review on genetics of ageing focused on two genomic markers of ageing: TL and chromosomal mosaicism (mosaic loss of chromosome Y (mLOY) in particular).

1.1 Human genome variation

Understanding of human genome variation lays the foundation for dissecting genetic pathophysiology of complex traits and diseases. High-throughput genotyping and sequencing technologies, with comprehensive curation of nearly all types of human DNA polymorphisms have driven the discovery of new disease-associated genes through genome-wide association studies (GWAS)[3]. Over the past decade, over 159,202 variant-trait associations have been reported, providing novel insights into genetic mechanisms underlying phenotypic variations in humans[4]. Despite these achievements, there are still significant challenges to overcome in order to better elucidate the relationship between genotypes and phenotypes. Rare and low

frequency variants (<1% and 1-5% minor allele frequency (MAF)[5], respectively) are incompletely characterised in most of the previous genotyping chip-based association studies due to genotyping quality and power limitations, yet they extensively outnumber the common variants (>5% MAF)[6,7]. Moreover, once a disease-associated locus has been identified, further characterisation of functional involvement of the locus requires extensive uses of both bioinformatic and experimental tools. Therefore, larger sample sizes, more comprehensive and accurate haplotype reference panels, and various gene prioritisation methods are required for a deeper understanding of genetic architecture of human phenotypes.

1.1.1 Structure of human genome

Human genome consists of nuclear and mitochondrial components, with the former storing the vast majority of genetic materials, encoded into ~3.2 billion pairs of deoxynucleotides (monomer units of deoxyribonucleic acid (DNA)) that form two anti-parallel strands configured into a "double helix" structure[8,9]. They are wrapped around nuclear proteins, forming 22 pairs of autosomal chromosomes and two sex chromosomes, X and Y. DNA sequence is defined as the order of four nucleotide bases, adenine (A), cytosine (C), guanine (G) and thymine (T). These sequences are organised into different functional elements, including exons, introns, non-coding RNA transcribed fragments and intergenic sequences, with 5% evolutionarily conserved among mammals and other vertebrates[10]. There are around 20-25 thousand protein coding genes based on current estimation[11] comprising ~1.2% of the human genome.

1.1.2 Genetic variation

Around 0.1% of DNA sequences differ between humans worldwide[12]. These differences in the DNA sequences are defined as genetic variation, which can occur at different scales, ranging from alterations in chromosomal quantity and structure (~3Mb or more), to small indels (< 1kb) and single nucleotide polymorphism (SNP, = 1bp)[13]. These can occur somatically in

various tissues, or germline cells which are capable of being transmitted to subsequent generations.

'Reference' human genome sequences have provided the basis for building a comprehensive landscape of genetic variants across whole spectrums of allele frequencies and types of variations[3]. In a landmark study from 2005, the International HapMap Project Consortium published their first map of common variation in human genome that contains more than one million SNPs obtained from 269 DNA samples from 4 populations (the Yoruba in Ibadan, Nigeria; Europeans in Utah, USA; Han Chinese and Japanese)[12]. Since then significant progress has been made with the Haplotype Reference Consortium (HRC) currently reporting over 39 million SNPs covering a large proportion of rare and low frequency variants in 32,488 samples with predominantly European ancestry[14]. These sources, with new and better design of genotyping chips, have largely expanded the number of SNPs identified in populations across different ancestries. More recent development in genome sequencing has improved identification of genetic variants with an even broader coverage of rare and low frequency variants and better accuracy, advancing genetic association studies into a new stage.

1.2 Genome-wide association study (GWAS)

GWAS represents a "hypothesis-free" approach to genetic discovery that systematically tests associations between genetic variants and continuous phenotypes, such as risk factors, or binary outcomes, such as disease endpoints, in cohort or case-control studies. This approach has proven extremely successful in robustly detecting associations between SNPs and diseases and quantifiable phenotypes, such as circulatory metabolite and protein levels, gene expression levels and cardio-metabolic risk factors[15–18]. Due to sample size expansion and new analytical method development, an increasing number of loci have been identified over the last decade[4,19]. However, as association does not necessarily indicate causation, for most of the identified genetic associations, causal variants within associated loci are unknown, and thus mechanisms of how these loci influence traits of interest are largely unveiled.

1.2.1 Rationale and basic principles

Experimental design of GWAS relies heavily on the principle of linkage disequilibrium (LD), i.e. non-independent associations between genetic alleles within a population. This largely occurs when alleles in close proximity are less likely to be separated during recombination and tend to be co-inherited as haplotype blocks. Large number of recombination events occur over generations, which shrinks haplotype blocks such that distances between variants tagging haplotype blocks extend over short distances (<0.1Mb) on average[20]. By assessing LD structures in reference genomes, carefully selected sets of variants on conventional genotyping arrays, with total numbers ranging from several thousands to millions and allele frequencies from rare to common, can cover the majority of human genomes, thereby reducing complexity and cost of assessing all SNPs genome-wide[21].

In addition, statistical imputation leveraging haplotypes estimated from fully sequenced reference panels can be used to predict untyped genotypes, help correct genotyping errors and facilitate cross-study association meta-analyses. The quality of genetic imputation heavily depends on the match of allele frequencies between genotyped tag variants and ungenotyped likely-causal variants[21,22]. Larger sample sizes of reference panels have improved imputation accuracy and coverage of rare and low-frequency variants, not only for European ancestry, such as panels curated by UK10K[7] and HRC[14], but also for diverse ancestries, including panels by 1000 Genomes[3] and International Haplotype Map Project Consortium[12,23].

1.2.2 Key considerations in GWAS

A primary consideration for GWAS is to determine whether significantly associated genetic variants are truly related to a phenotype of interest. Therefore a rigorous framework is required to distinguish causal from spurious signals across all stages of GWAS including extensive quality control (QC) checks, association testing methodology, and assessing robustness and validity of statistical inferences[24].

QC checks are important in reducing spurious associations, which include QC for both genetic and phenotypic data. I will describe QC procedures in more detail in subsequent

chapters but in brief, genetic QC includes checks for batch or study-centre effects, deviation from Hardy-Weinberg equilibrium[25], patterns of missingness, haplotype phasing and imputation of missing genotypes. Phenotype QC includes examination of outliers, checks for missingness, measurement errors and data distribution and normality, and subsequent transformation and standardisation of the data.

For association testing, various statistical methods have been developed[26]. For example, BOLT-LMM using a Bayesian mixed model association method is particularly designed to conducting biobank-scale GWAS. It was shown to deliver analyses in full UK biobank cohort in a few days on a single compute node, generating robust and powerful test statistics, while controlling for population structures[27]. For relatively smaller-scale GWAS, simple statistical models remain the most commonly used method of choice, such as logistic or linear regressions for binary or continuous outcomes, respectively. SNP genotypes or imputation probabilities are used as primary exposures, with one SNP tested each time with adjustments for potential confounding factors, such as age, sex, top principal components (PC) that represent an overall population structures, and other study-specific covariates. An additive mode of inheritance is most commonly assumed for each individual SNP, i.e. a heterozygote of disease-predisposing variants carries an intermediate risk between two homozygotes. Violation of this assumption, such as for SNPs with dominant and recessive effects, may lead to power loss.

Standard criteria to control false positive discovery while maintaining power is crucial for setting a balance between specificity and sensitivity of findings. Frequentist approaches are often used in GWAS, where statistical evidence of significance is measured by *p*-values – the probabilities of obtaining a value (in a population) that is the same or more extreme than the observed value (in a sample) when the null hypothesis (H_0, often referring to an zero effect size of a genetic variant on a trait of interest) is true. In GWAS, as millions of SNPs being tested, keeping the significance at the conventional *p*-value threshold of 0.05 can lead to a large number of false positive association results (type 1 error: wrongly rejecting H_0)[28]. Bonferroni correction, assuming each individual SNP as an independent test, overlooks correlations between SNPs in LD, and thus can be overly conservative. By taking into account of LD of variants across the genome, a standard *p*-value of 5×10^{-8} was proposed and is currently widely accepted as the genome-wide significance threshold for studies in European populations regardless of imputation density[29–32]. However, quantifying *p*-value alone is

insufficient to assess how likely a given SNP is truly associated with a phenotype. This is because in very low powered scenarios (e.g. due to either low SNP MAFs or small sample sizes), small *p*-values which may seem to offer strong evidence against the null hypothesis (H_0) also indicate similar less likelihood under H_1 (the alternative hypothesis)[33].

Alternative ways for defining significance thresholds have also been proposed. One is based on permutation tests that use actual data to empirically evaluate probabilities of observing a more extreme result by chance[34]. Another is to use false discovery rate (FDR), by which the limitation of *p*-value is offset by first ranking *p*-values in an ascending order and then correcting them by their relative ranking positions[28]. The quantile–quantile (QQ) plot of log *p*-values can be used to visually interpret results with regards to their overall significance levels, wherein negatively ranked log *p*-values are plotted against corresponding null expectations (*i/(n + 1)* for the i^{th} smallest *p*-value of *n* tests assuming H_0 is true for all tests)[28,35].

Moreover, Bayesian methods provide an alternative way of assessing significance of association results[36]. Using these methods, a posterior probability of association (PPA) is calculated for each SNP, jointly determined by evidence from observation (Bayes factor: the ratio of probabilities under H_1 and H_0) and prior knowledge (prior probability (π)), and thus is insensitive to statistical power and number of tests performed[33,37]. The value of π can vary across SNPs, depending on MAFs or other bioinformatic annotations of SNPs; or be set to a constant value for all SNPs, indicating an overall proportion of SNPs being truly associated with traits of interest. However, Bayesian methods have not been used as commonly as the frequentist approaches, possibly due to a greater computational demand and an inconsistent, subjective pre-specification of π. Recent methodological developments, such as the BOLT-LMM method, showing improved computational efficiency and more flexible specification of π, has become increasingly useful for GWAS practitioners, especially in biobank-scale studies[27].

1.2.3 General results

The first GWAS was performed in 2005, which studied age-related macular degeneration in a case-control cohort of Southeast Asians. It examined 97,824 associations of autosomal SNPs in 226 participants[38]. The first large scale consortium effort by the Wellcome Trust Case

Control Consortium in 2007 examined associations with 7 common complex diseases, each consisting of ~2,000 cases and a shared set of ~3,000 controls[39]. This was widely regarded as a landmark GWAS in the field due to its substantial scale at the time and that it marked the beginning of an exponential rise in GWAS[15,40]. Since then, thousands of genetic loci have been identified to be associated with hundreds of complex traits, ranging from common diseases, biomarkers, brain imaging and anthropometric phenotypes, gene expression and protein levels, and sociobehavioural traits. As of September 2019, the GWAS Catalog has recorded 4,220 research papers (PubMed ID) that contained 7,661 study-specific traits (Study Accessions) across 4,669 unique human phenotypes that can be mapped to 2,608 Experimental Factor Ontology traits. The average number of significant associations identified with each study-specific trait is 20. However, the thresholds of significance varied, which on average were set as the *p*-value of $1X10^{-6}$, even though ~66% of reported associations have reached the canonical level of genome-wide significance threshold (*p*-value=$5X10^{-8}$). The sample sizes also varied from several hundreds to over a million individuals with single or mixed ancestry backgrounds. Although only ~47% of the SNP-trait associations reported have been tested in independent data sets and proven to be replicable, evidence has proven that GWAS results in general are highly reproducible, even across ethnicities, on condition that LD patterns underlying causal variants are similar between populations[41].

1.2.4 Heritability and genetic architecture

Heritability measures proportions of total phenotypic variations attributable to genetics. This is classically estimated from empirical data with informed relationships of individuals (such as offspring-parents, siblings and twin pairs). These classic estimates of heritability can be categorised into two groups: narrow- and broad-sense heritability. The latter expands on the former by also including interactions between alleles at the same or different loci[42]. However, estimates from these classical approaches are susceptible to biases originating from assortative mating and natural selection, and their accuracy depends on sampling variation, which is jointly determined by sample size and pedigree structure[42]. GWAS provided a novel way of estimating the narrow-sense heritability based on co-segregation of alleles tagged by genotyped and imputed SNPs across human genomes, referred to as the "SNP heritability"[21]. In contrast to the classical methods, the SNP heritability is estimated based on genetic

relationship matrices, which are calculated in large-scale population cohorts that are often unrelated and do not necessarily contain pedigree information[21,27]. Therefore, it can avoid the conventional biases and improve accuracy with increasingly larger sample sizes.

Common complex diseases, unlike Mendelian disorders, are often driven by multiple genetic variants, with each explaining a small proportion of the total heritability. Current technologies using SNP-array based genotyping platforms restrict findings to mostly common variants, whereas how rare variants contribute to heritability estimation is largely unexplored. With the emergence of large-scale next-generation sequencing, we can investigate whether and how much rare variants increase the total variability explained, thereby facilitating discovery of overall genetic architecture of diseases.

Genetic architecture characterises the impact of individual variants on a broad sense of phenotypic variability[43]. It describes heritability at a finer resolution than the overall SNP heritability, which includes total numbers of variants associated, correlations between effect sizes and allele frequencies of the variants, and potential interactions between variants and between variants and non-genetic environment[44,45]. Genetic architecture varies between different phenotypes and diseases in terms of total numbers of variants associated and variants' allele frequencies and effect sizes. For example, the two types of diabetes mellitus are both highly heritable and polygenically-driven, yet with distinct genetic architecture. Type 1 diabetes is mainly driven by a few large-effect variants with relatively low-frequencies on average[46,47], whereas type 2 diabetes (T2D) is cumulatively driven by a large number of common variants with small effect sizes[44]. In addition to complex diseases, intermediate traits, such as anthropometric traits and biomarkers, also demonstrate different genetic architecture[43,48]. For example, there are 47 independent variants associated with plasma glycine levels, exhibiting a broad range of effect sizes and allele frequencies, whereas only 6 with 25-hydroxyvitamin D, most of which are common variants with small effect sizes[49,50], even though the sample sizes and SNP imputation densities are comparable between the two GWAS.

1.2.5 Main challenge and new approaches

A main challenge for GWAS is to pinpoint possible causal variants and genes among association signals. GWAS results typically cannot provide direct evidence to pinpoint causal

variants and genes because (a) many of the significantly associated variants are in non-coding or intergenic regions with no known functional implications, and (b) the presence of LD leads to multiple variants statistically showing similar association strengths to causal variants.

Rare variants with functionally detrimental consequences on gene products, such as missense and nonsense mutations, if showing robust association signals, are often causal due to natural selection against functional mutations in surviving essential genes. However, associations with rare variants are difficult to identify due to limitations in genotyping and imputation methods and reduced power of association tests for rare variants. Recently, rapid advances in sequencing technologies have enabled a more complete and accurate identification of low-frequency and rare variants, and with rare variants-oriented GWAS methods, these can potentially enhance the capability of GWAS to detect rare variant associations. For example, the gene-based burden tests[21] and variance component tests[51,52], in which effects of rare variants are combined through aggregation algorithms, can boost power of association tests for rare variants. Scalable approaches, such as SAIGE-GENE, can handle large sample sizes at biobank-scale, further improving statistical power in detecting rare variant associations[53].

Post-GWAS annotation can also facilitate selection of likely causal variants and genes. Using multiple bioinformatic sources, such as VEP (Variant Effect Predictor)[54], UCSC human genome[55], ENCODE (ENCyclopedia Of DNA Elements)[56], and Exome Aggregation Consortium[57] databases, a variety of information can be obtained for variants in their potential roles in gene transcriptional regulation and disease pathogenicity and their evolutionary conservation levels. For example, variants that play roles in gene transcription, such as those located within histone modification markers or transcription factor binding sites, are more likely to be causal, as well as variants that are more conserved across species, and with aetiological relevance to diseases. Such functional annotations of variants can be integrated into fine-mapping approaches, for example, through weighting and adjusting prior probabilities in Bayesian models[58]. Moreover, linking variants with gene expression quantitative trait loci (eQTL) can also help infer causality. Besides simple search of individual SNPs across relevant databases, such as Genotype-Tissue Expression (GTEx) database[59], integrative methods have recently been developed, such as transcriptome-wide association (TWA) and statistical colocalization methods. The former imputes gene expression levels via incorporating eQTLs with individuals' genetic profiles and correlates the imputed gene expressions to traits of interest[60–62]; the

latter tests probabilities of gene expressions sharing the same causal variants with the traits of interest[63,64]. These methods have demonstrated extreme power in prioritising likely causal variants and genes, underpinning methodological development in the future post-GWAS era.

1.2.6 Statistical application of GWAS results

Sharing of GWAS summary statistics (effect sizes of millions of SNPs and their standard errors or estimates of equivalent parameters for associations with different traits) have benefited the entire field of population genetics, because it enables comparison and integration of results across studies as well as stimulating method development. This has facilitated discovery of novel disease-predisposing loci in larger meta-analyses, improvement of accuracy in estimating SNP heritability, more comprehensive characterisation of genetic architecture and 'in silico' investigation of specificity and pleiotropy for individual SNP effects[21]. Moreover, owing to this tremendous data resource, various statistical approaches have been developed which uses summary statistics to infer causalities and predict disease risks, thereby deepening our understanding in disease aetiologies and facilitating clinical utilities.

1.2.6.1 Mendelian randomisation (MR)

Evidence from double-blinded, randomised controlled trails (RCTs) are considered the gold standard for causal inference, because study participants are selected with balanced distribution of known and unknown confounders, yet randomly assigned to intervention and control groups to minimise selection bias[65]. RCTs are often infeasible, as they are extremely expensive, difficult and time consuming, and may not be ethically appropriate, depending on exposures of question. In contrast, observational studies in epidemiology are prone to confounding and reverse causation, as participants in such study cohorts are randomly recruited with minimal controls for confounding factors and undetermined time sequence of exposures and outcomes. Hence, statistically significant associations found and reported in such studies do not allow inference about causality[66].

MR studies have been proposed as a natural analogue of RCTs, in which randomly allocated genetic alleles are used as instrumental variables that mimic randomised groups of

RCTs, i.e. absence or presence of genetic alleles mimic control and intervention groups in RCTs, respectively. Similar to RCTs, MR aims to investigate putative causal effects of risk factors on outcomes while minimising influence of confounders[67]. Because genetic variants are fixed at conception and randomly assigned during meiosis, therefore using them as proxies for exposures of interest can overcome the two major limits of observational association studies: unmeasured confounding and reserve causation. Three key assumptions must be satisfied for MR analyses. The genetic instruments must (1) be strongly associated with risk factors of interest (relevance assumption), (2) not associated with any confounding factors (independence assumption) and (3) influence outcomes exclusively through pathways that are mediated via the risk factors (exclusion restriction assumption)[67–69].

To ensure the relevance assumption, genetic instruments are often selected based on evidence obtained from published GWAS. Ideally data sets from which genetic instruments are selected should be different from but within the same underlying populations as those for causal association inferences, the so-called "two-sample MR framework"[70,71]. One-sample MR, where estimates of associations with risk factors and outcomes are derived in the same data sets, may suffer from the 'Winner's curse'[72]. Instrumental variables are often constructed from multiple genetic variants known to be associated at the level of genome-wide significance. Advantages of using multiple variants are that (1) they collectively explain more variability of risk factors than using single variants, and thus are statistically more powerful, and (2) potential pleotropic effects are diluted, although unlikely to be eliminated entirely[73]. Variants at independent loci are most commonly used, whilst correlated variants can also be used with adaptive methods that incorporate variance-covariance matrices[69]. Including additional variants, even if correlated to a certain degree, can improve accuracy of MR analyses, however, at the cost of exacerbating 'weak instrument bias' due to the overfitting problem. Similarly, applying a lower association threshold to variants can result in a greater number of variants selected, and thus increase power but also at an increased risk of overfitting.

Unbalanced pleiotropy of genetic variants is a major concern of violation of independence and exclusion restriction assumptions. Although these assumptions are not fully testable, a systematic investigation of associations of each genetic instrument with a broad spectrum of phenotypes can help verify these assumptions. These tests are analogous

to checking equal distributions of potential measured confounders between treatment and control groups in RCTs[67].

Owing to an increasing number of summary association estimates generated via GWAS, the MR methods have been expanded with enhanced statistical power and allowing for relaxation of some of the MR assumptions, including MR-Egger regression and (weighted) median MR. The MR-Egger regression, developed on the basis of the Egger's test - a test that assesses small study bias in meta-analysis, provides a valid method of detecting directional (unbalanced) pleiotropy via testing the hypothesis of the Egger's intercept being equal to 0. In scenarios where the assumption of exclusion restriction is violated due to unbalanced pleiotropy, MR-Egger can be applied to generate consistent causal estimates under the InSIDE (Instrument Strength Independent of Direct Effect) assumption - association strengths of genetic variants with exposures are independent of direct effects of genetic variants on outcomes[68]. Similarly, the (weighted) median MR method can tolerate up to 50% invalid instruments, because the method takes median instead of mean of causal ratio estimates[74].

1.2.6.2 Polygenic risk score (PRS)

Genetic risk profiling, considered as an early measurable predictor of disease risk, has demonstrated crucial values in disease risk prediction and prevention. Previous clinical utilities of genetic risk profiling have largely focused on rare functional mutations embedded within causal genes for rare monogenic diseases. For complex polygenic diseases, the genetic risk profiling can be summarised and assessed using PRS[75–79].

Construction of PRS is similar to that of an instrumental variable (i.e. allele score) in MR, both of which are calculated as a weighted sum of genetic risk alleles of all associated regions across individual human genomes, where the weights, defined as association effect sizes, are obtained from GWAS[80]. However, considerations of what and how many genetic variants are included into PRS may be different from those into an instrumental variable in MR, as the former focuses on prediction of disease risks, whereas the latter on causation between risk factors and disease outcomes. A threshold of genome-wide significance is most commonly used when determining the set of genetic variants to be incorporated into PRS construction, but a lower threshold may also be used, thereby increasing total variability of diseases explained and predictability of PRS, often at the cost of reduced generalizability in clinical applications (for example, to individuals of different ethnicity backgrounds)[81–83]. An optimal

threshold depends on sample sizes and genetic architecture underlying traits of interest. Inclusion of additional variants that are below genome-wide significance may harm the performance of prediction models (often assessed by areas under the receiver-operator curve (AUC)), especially when the variants are obtained from insufficiently powered GWAS[84]. In contrast, for certain late-onset diseases, for which large-scale, well-powered GWAS have been undertaken, such as T2D[44], coronary artery disease (CAD)[85] and Alzheimer's disease (AD)[86], inclusion of more modestly associated variants can improve prediction accuracy because their genetic determinants have been demonstrated to mainly consist of common variants of small effect sizes, with little or no evidence showing rare variants of large effect sizes are involved in their disease aetiologies[80].

Clinical utilities of PRS are still under debate, especially considering common conventional non-genetic risk factors, such as age, gender and behavioural and environmental risk factors explain the majority of phenotypic variance[89]. Moreover, for many diseases, heritability estimates attributable to PRS are still limited, even with increasingly larger numbers of disease-susceptibility genetic loci identified, (e.g. ~20% by over 400 loci for T2D)[44,88]. Applying the most recent PRS of T2D to participants in UK biobank produced an AUC of ~65%[87], which, although substantially increased previous prediction performance[89–91], was still worse than simply using conventional T2D risk factors, such as age, gender, family history and biomarker and adiposity levels. However, the relatively small proportion of heritability explained may not necessarily restrict possible utilities of PRS in disease prediction. For example, mutations within *Breast Cancer Susceptibility Protein Type* (*BRCA)1* and *BRCA2* genes are rare in general populations (<<1%), and the number of incident cases of breast cancer that carry these mutations is small (~5% of all cases), so that heritability explained by these mutations is also limited. However, the relative risk of developing breast cancer due to these mutations is large (4-5 folds)[80]. Therefore, using such highly penetrant pathogenic mutations in certain diseases can help clinical practitioners take earlier interventions to prevent disease occurrence.

PRS can be helpful in identifying a subset of population who are at extreme tiers of risks for common complex diseases, and the subset may vastly outnumber extreme phenotypes due to single pathological gene mutations. For example, using PRS alone identified 8% of a general population who were at a 3-fold higher risk of CAD, whereas only 0.4% of the same population carried familial hypercholesterolemia mutations that exert similar risk effects on

CAD[92]. Overall PRS-informed population stratification and individual risk assessment for certain diseases can facilitate personalised medicine, and better inform clinical communities in making decisions in disease prediction and prevention[80,84].

1.3 Genetics of ageing

The proportion of elderly populations is growing rapidly worldwide and it is expected that proportion of those aged over 60 years will be doubled over the next three decades[93]. Ageing is often associated with a progressive functional decline that involves a variety of physiological and psychological changes that impair organ functionality and rejuvenescence, memory and cognition, and overall physical performance and intellectual ability[2]. Risks of many common complex diseases, including different cardio-metabolic diseases, cancers and neurodegenerative disorders, increase with age, resulting in a profound reduction in life quality of older populations and a huge burden on social and health care systems[76,93–95].

There is a substantial heterogeneity in life expectancy between individuals, which is largely driven by both environmental and genetic factors[2]. Based on twin studies, the heritability of human lifespan has been estimated to range between 20-30%, although estimates differ between studies[96,97]. A recent study has shown that heritability estimates of human longevity may be inflated due to assortative mating and the true heritability may be less than 10%[98]. Understanding genetic mechanisms that regulate ageing may provide insights into aetiologies of age-related diseases, and ultimately lead to novel approaches to reduce age-related morbidity and mortality rates and improve quality of life[1].

From experimental studies in animal models, genes that play fundamental roles in conserved pathways have shown lifespan-altering capacities[99]. These so-called gerontogenes (gene expressions negatively associated with longevity) or longevity-assurance-genes (gene expressions positively related to longevity) are often pleiotropic, as they are involved in different pathways that regulate hallmarks of ageing[100]. These hallmarks have been categorised into nine main interconnected domains: genome instability, TL attrition, epigenetic alterations, loss of protein homeostasis, dysregulated nutrient sensing,

mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication (Figure 1.1)[1,2].

Figure 1.1 Hallmarks of ageing.
The nine hallmarks of ageing are interconnected with each other through shared molecular components and pathways. They interact with environmental signals to jointly determine trajectories of lifespan. The figure was adapted from López-Otín et al[2].

1.3.1 Genetic studies in animal models - evolutionarily conserved pathways

Fundamental mechanisms that regulate ageing and related phenotypes converge onto several classic signalling pathways and gene sets, including the well-characterised mechanistic target of rapamycin (mTOR), insulin/insulin-like growth factor (IGF), and adenosine monophosphate-activated protein kinase (AMPK) pathways, and sirtuin

deacetylases[1,101]. Under normal circumstances, genes implicated within these pathways regulate growth and metabolism, while in extreme conditions (e.g. dietary restriction, oxidative stress and extremely high or low ambient temperatures), their functions adapt accordingly to protect organisms from those environmental stresses, resulting in a shift of physiological states towards a standstill that facilitates global maintenance and stability, and thus slowing down growth and ageing processes[101]. Each condition involves specific proteins, but the multifaceted functionalities of these proteins imply their involvements in multiple pathways, through which interconnections between regulatory pathways of ageing are established. Because of this network of ageing regulatory pathways, pharmacological or genetic interventions that target individual gene products are sufficient to slow down ageing, even though ageing itself is a complex phenotype that is subjected to numerous regulatory mechanisms. Such experimental hypothesis underscores the latent potential of extending lifespan via drugs that target single molecular candidates. For example, inhibitors of the mTOR pathway (such as rapamycin and several derivative compounds) have been clinically approved or under development for a variety of clinical uses, including therapies for post-transplantation and several types of cancers[102,103]. Metformin, an anti-diabetic drug that activates AMPK, can manipulate ageing phenotypes and healthspan in lower organisms through multiple mechanisms, including downregulation of the insulin/IGF signalling and the mTOR signalling pathways[101].

Insulin/IGF signalling is the first pathway that has been shown to influence longevity in *C. elegans*, wherein loss-of-function mutations in an orthologue (abnormal dauer formation-2, *daf-2*) of the *IGF receptor* (*IGFR*) gene left worms arrested in juvenile forms with more than doubled healthy lifespan[104,105]. In humans, lower levels of IGF-1 hormone[106], or loss-of-function mutations in the *IGF1/2R* genes[107] or in the related transcriptional factor gene, *Forkhead box O 3* (*FOXO3*)[108–112], have been associated with longer survival in centenarians. Further, growth hormone (GH) has been negatively correlated to longevity in both mice and humans through mechanisms that were independent of IGF1[1,113]. A homozygote deletion variant in the exon 3 of the *GH receptor* gene has been positively associated with older age in long-lived human cohorts[114].

Most genes identified in animal studies lack replication in human genetic studies, which will be further discussed in section 1.3.2. Nevertheless, these candidate genes and signalling pathways primarily found in animal models provide a potential catalogue of drug targets for

ageing and age-related diseases, yet validation of their roles in humans is still the key to facilitate further investigation for any pharmaceutical potentials.

1.3.2 Human genetics of aging

The complexity of human ageing lies in joint effects of a variety of risk factors, including genetic, behavioural and environmental risk factors, such that the extent to which genetic variation affects human lifespan is under debate. Thus far, most of the candidate genes that manipulate lifespan have been identified through gene screening in non-vertebrate model organisms, with only a handful of them supported by evidence from human population studies[115]. Many of these genes overlap with genes associated with age-related complex disorders, revealing shared pathophysiological pathways between ageing and diseases[97].

1.3.2.1 Candidate gene studies

Few genes have been consistently reported to be associated with human longevity and lifespan across different studies, including *Apolipoprotein E* (*APOE*), *Cholinergic Receptor Nicotinic* (*CHRN) Alpha 3/5 Subunit*, *Human Leukocyte Antigen* (*HLA*)*-DQA1/DRB1*, *lipoprotein A* (*LPA*), *FOXO3*, *IGF1/2R*, *SIRTURIN 3* and *Interleukin 6* (*IL6*)[1,108,109,111,112,116–120]. Although hundreds of genes have been suggested from animal model experiments, they failed to replicate in human studies, possibly due to the complexity of human ageing, which restricts genetic discoveries to be highly context-dependent. Environmental factors, such as geographical and anthropological segregation, socioeconomic status and education, can bias genetic association analyses and affect interpretation of novel findings[116]. Therefore, generalisability of results is often limited and possibly under specific study populations and times. Moreover, the general lack of reproducibility can also be due to different study designs, measurements of longevity outcomes, age stratifications and power[121], which will be further discussed in section 1.3.2.2.

1.3.2.1.1 Informed by model organisms

Candidate genes selected for human association studies have mainly been based on experimental evidence from animal models. They are involved in regulation of different

ageing hallmarks or conventional risk factors of age-related diseases. For example, *IGF1/2R* and *FOXO3* genes are involved in the nutrient sensing pathway, and their orthologues have been shown to play important roles in model animals (section 1.3.1). *HLA-DQA1/DRB1* and *IL6* genes are involved in the immune function and inflammation; *CHRNA3/5* locus, encoding various subunits of nicotinic acetylcholine receptors, is associated with smoking-related behaviours and diseases, and smoking is a well-established strong cause of premature death; *APOE*, encoding a receptor-binding ligand on the surface of multiple lipoprotein particles, and *LPA*, encoding a constitutive protein component of lipoprotein(a), have been linked to an array of age-related diseases and relevant risk factors, including AD, CVD and total and low-density lipoprotein cholesterol and triglycerides[119,120]. Together, these findings have demonstrated that candidate genes emerging from animal studies can inform relevant epidemiological studies in human populations, but replication may need larger cohorts with diverse ethnic backgrounds and careful consideration of gene-gene and gene-environment interactions.

1.3.2.1.2 Informed by rare diseases

A group of rare premature ageing diseases (e.g. Hutchinson-Gilford progeria syndrome (HGPS) or Werner syndrome (WS)), known as progeria, with clinical presentation of dramatic and premature ageing, i.e. early appearance of phenotypes generally associated with ageing, such as atherosclerosis, osteoporosis, greying and loss of hair, skin ulcers and occurrence of multiple rare cancers[1]. Studies of genetic aetiology of progeria can help to understand regulatory mechanisms of normal aging. For example, nuclear aberrations, including altered histone modification patterns and increased DNA damage, were found in cells from HGPS patients, as well as in skin fibroblasts from older individuals in general populations[122]. Donor cells from older individuals showed a remarkable change in the nuclear location of Lamin A/C (LMNA): changing from nuclear lamina at nucleoplasmic side of inner nuclear membrane to nuclear periphery, thereby disrupting integrity of nuclear membranes, leading to dysregulated cell cycle and gene transcription; similar dislocation of LMNA was also found in cells from progeria patients that possess truncated LMNA isoforms[122,123]. Moreover, the *WRN* (*Werner Syndrome RecQ Like Helicase*) gene that causes WS, encodes an ATP-dependent helicase that functions in DNA replication, transcription, repair and recombination, and

telomere maintenance. This supports genome instability as a hallmark of both, premature ageing disorders and normal ageing processes[124,125].

1.3.2.2 GWAS

GWAS provides a hypothesis-free method of characterising genetic architecture of human ageing. It has the potential of identifying novel genes and perhaps human-specific mechanisms of regulating age-related physiological changes. Case-control studies, in which people who live exceptionally longer are included as cases and compared to those with normal lifespan have been conducted[111,126,127]. This simple design is often restricted due to inaccessibility of control samples who are born in the same period of time (early 90's) as the long-lived cases but die at younger ages. Using alternative controls from later generations may introduce bias towards the null due to environmental variation and secular trends in factors affecting longevity across generations[126–128]. To minimise selection bias, a prospective birth cohort design is employed with participants followed up from birth and outcomes defined as continuous measures of lifespan (e.g. time to death or to the first incidence of fatal diseases)[129]. More recently, a cross-generational design has been developed, in which parental age at death was regressed on genetic variants obtained from offspring. The rationale behind this method is that a) half of genetic materials are shared between parents and offspring, b) there is a positive correlation between general health states of middle-aged individuals and their parents' ages, i.e. people who have longer-lived parents are generally healthier[130]. This study design can increase effective sample sizes, given that the number of death events among parents is more than doubled than that in participants (offspring), especially in large population cohorts where the majority of participants are recruited in their middle ages and hence may have aged or deceased parents[120]. For example, UK biobank is a such resource, in which half a million participants were recruited at ages 40-69 years and over 60% of their parents were recorded as dead at baseline[120].

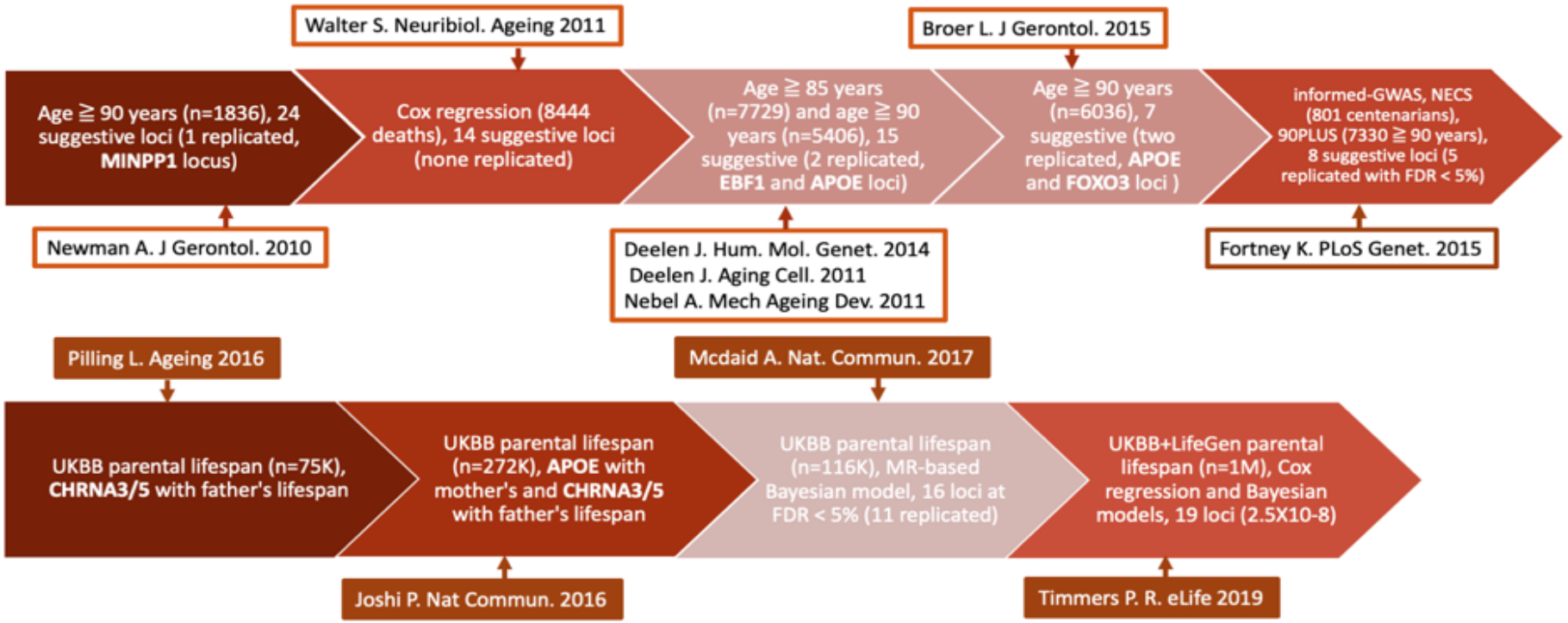
New analytical approaches can also facilitate novel gene identification. For example, a concept of the "informed GWAS" that integrates prior knowledge of age-related disease loci has been realised by different approaches, one based on a multivariate MR model with Bayesian priors[97] and another using a centenarian enrichment approach[131] (Figure 1.2). Both of these approaches are successful in identifying novel genes associated with human lifespan,

indicating ageing and age-related diseases are interconnected through shared genetic determinants.

However, despite a large increase in statistical power in recent GWAS on human longevity and lifespan (sample sizes increased from several thousands to one million, Figure 1.2), only a few loci have consistently shown genome-wide significant association signals across studies. These include not only *APOE* and *FOXO3* genes that have consistently been reported in both animal and candidate gene studies, but also previously unidentified loci with potential roles in various hallmarks of ageing. For example, the *CDKN* (*Cyclin Dependent Kinase Inhibitor*) *2A/B* genes, consistently reported by several large-scale, parental longevity studies[118–120,132], is involved in regulation of cell cycle and cellular senescence. Dysfunction of this protein can lead to stem cell exhaustion and increased risks of many age-related diseases, including cardiometabolic disorders and cancers[97,118,121]. Similarly, the *MAGI3* (*Membrane-Associated Guanylate Kinase Inverted 3*) gene, encoding a membrane-associated guanylate kinase that acts as a scaffolding protein at cell-cell junctions, thereby facilitating intercellular communications, has been associated with autoimmune diseases and lifespan[118,133]. *LPA* and *LDLR*, together with *APOE*, the first well-documented human longevity locus, are all involved in transport and metabolism of lipoprotein particles, thereby influencing levels of lipid risk factors for cardiometabolic diseases[118]. These show that genes with functional involvements in hallmarks of ageing may be associated human lifespan and can be identified in general population studies.

Although recent GWAS have expanded longevity loci from one (*APOE*) to more than a dozen with sample sizes increased from several thousands to more than 1 million individuals, power is still a limiting factor due, at least partly, to relatively small numbers of death events in large-scale general population cohorts, as which are often in short of follow-up time. Moreover, not all deaths are attributed to heritable age-related diseases, other factors may also influence individual lifespans and such confounding factors may be overlooked in GWAS, causing lifespan-associated loci rather context-dependent, and thus reducing power in meta-analyses[132]. In order to detect more novel and robustly associated loci, larger sample sizes, increased coverage of rare variants and more advanced mathematical approaches may help.

Figure 1.2 A timeline showing evolution of GWAS on human longevity and lifespan.



1.4 Genomic markers of ageing

While a large body of research has focused on identification of biomarkers of age-related diseases, assessing their values for prediction of disease risk and response to clinical interventions, no consensus exists for biomarkers of 'biological age' – indicators that can capture and measure ageing process as opposed to actual chronological age[134]. Several criteria have been proposed as prerequisites for such markers, including they should (1) demonstrate a continuous change along with age and allow for longitudinal monitoring, (2) reflect physical and cognitive decline associated with ageing, (3) and predict age-related multi-morbidities and mortality better than chronological age[121]. Recent studies have highlighted the importance of genomic changes that occur while we age, including TL shortening and chromosomal loss, which have been associated with cancers and possibly also other age-related diseases[1,2,135]. In this section, I will give a comprehensive literature review on two markers of genomic ageing, TL and mosaic chromosomal loss, mLOY in particular, corresponding to telomere shortening and genome instability hallmarks of ageing (Figure 1.1). Other markers of biological age (e.g. circulatory small molecule metabolites, blood pressure and frailty phenotypes), capturing other aspects of pathophysiological changes during ageing, are involved in the other hallmarks of ageing, which will not be covered in this section. For example, insulin and blood lipid levels indicate nutrient sensing, IL-6 and high-sensitivity C-reactive protein are inflammation markers involved in cell senescence, accumulation of reactive oxygen species and peroxidised lipids reflect mitochondrial dysfunction, and alterations in DNA methylation and histone modifications consist of epigenetic changes along with ageing[121,136–140]. The markers of genomic ageing, in combination with other well-established markers of biological age can facilitate more comprehensive characterisation of ageing-related physical and mental decline at a molecular level and improve early prediction of age-related diseases in elder populations.

1.4.1 TL

1.4.1.1 Definition, structure and function of telomeres

Telomeres are DNA-protein complexes found at the end of eukaryotic chromosomes, which serve to maintain genomic stability and determine cellular lifespan[141,142]. Telomeric DNA consists of a long tract (10-15 kb) of double stranded TTAGGG repeats with a guanidine-rich single stranded overhang at the 3' end[143]. Protein complexes, including the SHELTERIN complex [Telomeric Repeat Binding Factor (TERF)1, TERF2, Protection of Telomeres Protein 1 (POT1), Ras-Related Protein Rap-1 (RAP1), TERF1-Interacting Nuclear Factor 2 (TIN2) and Tripeptidyl Peptidase 1 (TPP1)] and the CST complex [Oligonucleotide/Oligosaccharide-Binding Fold-Containing Protein 1 (OBFC1), CST Telomere Replication Complex Component 1 (CTC1) and TEN1 Subunit Of CST Complex (TEN1)] along with DNA helicases, such as Regulator Of Telomere Elongation Helicase 1 (RTEL1), bind telomeres and regulate TL and telomere structure[144–146]. Together telomeres form a highly compact chromosomal configuration, protecting telomeric DNA from uncontrolled elongation or being recognized as double-stranded breaks that trigger DNA damage responses (DDR)[147]. Telomeres shorten with cellular divisions due to the end replication problem and once critical lengths are reached cells enter replicative senescence[142]. In some cell types, such as stem and germ line progenitor cells, TL is maintained by the enzyme telomerase, a ribonucleoprotein containing the RNA template [Telomerase RNA Component (TERC)], the enzymatic protein [Telomeric Reverse Transcriptase (TERT)] and accessory proteins, including Dyskerin Pseudouridine Synthase 1 (DKC1) and several nucleolar proteins[148].

1.4.1.2 Germline genetic variants associated with TL

TL has previously been estimated to be highly heritable (h^2=0.70 (95%CI=0.64-0.76)) according to a meta-analysis of family/twin-based cohort studies, although heritability estimates varied between studies, ranging from 0.34[149] to 0.82[150], likely due to differences in methods used to measure TL and study designs and cohorts (twin-, non-twin sibling pairs- or multi-generational relatives-based cohorts). Overall, these have suggested a strong genetic component underlying TL regulation.

GWAS performed to date have identified 10 genes to be associated with TL measures in leukocytes, including 6 known biologically relevant ones. Among these candidate genes, three of them, *TERC*, *TERT* and *Nuclear Assembly Factor 1* (*NAF1*) in 3q26, 5p15.33 and 4q32.2, respectively, encode RNA/protein products that are involved in telomerase ribonucleoprotein

complex assembly; two genes, *OBFC1* (also known as *STN1*, 10q24.33) and *CTC1* (17p13.1) encode protein members of the heterotrimeric CST (CTC1, STN1 and TEN1) complex, which regulate telomerase activity by controlling accessibility of telomerase to telomeric DNA substrates[96,151–154]. Mutations in the *CTC1* gene, disrupting configuration of the CST complex, were causally linked to Dyskeratosis congenita and Coats Plus syndrome[96]. *RTEL1* (20q13.3), encoding a DNA helicase that facilitates structural unwinding of telomeric DNA sequences, also plays an essential role in telomeric DNA replication. Besides *RTEL1*, genes encoding other members of the helicase protein family, such as *RecQ*, have been reported to be involved in maintaining telomere homeostasis[155]. Disrupted functions of these helicases have been associated with telomere syndromes, such as the Werner and Bloom syndrome. Additionally, three loci, including *Acylphosphatase 2* (*ACYP2*)[151], *PX Domain Containing Serine/Threonine Kinase Like* (*PXK*)[156] and *DEAH-Box Helicase 35* (*DHX35*)[157,158], have been reported once, but not replicated in other independent studies, and thus their biological relevance to TL regulation remain uncertain. Finally, the *Zinc Finger Protein* (*ZNF)208/ZNF257/ZNF676* (19p12) region has been reported to be associated with LTL by several studies, but biological mechanisms remain to be established[96,151].

1.4.1.3 Non-genetic risk factors associated with TL

It has been widely acknowledged that shorter TL is associated with older age[159]. Other factors have also been associated with shorter TL, such as male sex, increased adiposity[160–162], smoking[163,164], alcohol consumption[165], reduced physical activity[166]. Much of the evidence comes from observational studies and which factors are causally associated with TL or subject to bias, reverse causation and/or confounding is unknown. Moreover, results reported by prospective cohort studies have been inconsistent or contradictory, possibly due to differences in methods of DNA extraction and TL measurement, power, study design and sampling, and statistical approaches used[162].

A large meta-analysis that collected 63 observational studies has reported that 38 of them found significant inverse correlations between LTL and adiposity levels, however, with extremely large between-study heterogeneity (I^2=99%)[167]. The large heterogeneity has been suggested to be possibly due to differences in statistical methods used and diverse study populations included, but not due to differences in methods used for TL quantification, stratifications on age or obesity, tissue types or countries of origin[167].

Other modifiable factors have been reported to be associated with LTL, yet only by one or two studies, which may also suffer from similar study limitations as described above. These included serum leptin levels[160], self-reported physical activity levels in leisure time[166], sedentary lifestyle[164], educational attainment[168], violent exposures[169,170], paternal age at children's births and paternal lifespan[171].

Therefore, the degree to which TL is causally influenced by modifiable factors such as obesity is uncertain. MR with genetic instrumental variables may help to elucidate such causality questions. GWAS on various modifiable factors have identified genetic variants that are robustly and specifically associated with those factors, and these genetic variants can serve as good instruments, with which causalities can be investigated under the MR framework (section 1.2.6.1).

1.4.1.4 Clinical consequences of dysregulated TL

In healthy somatic cells, shortened TL functions as a mitotic clock that prohibits indefinite proliferation of cells, thereby avoiding accumulation of carcinogenic DNA mutations[172]. However, extremely shortened TL leads to accelerated aging, and premature ageing syndromes, commonly characterised with impaired capacity for tissue renewal and stem cell exhaustion[173]. Dysregulated TL has been causally linked to both rare monogenic and common polygenic diseases, the former based on identification of rare pathogenic mutations in rare diseases, whereas the latter by using genetic approaches to assess causality or observational inferences from large prospective cohort studies. Therefore, maintaining telomere homeostasis is essential for healthy lifespan.

1.4.1.4.1 Rare diseases (telomeropathies)

Telomeropathies are a cluster of rare monogenic diseases due to disrupted telomere homeostasis. These can be categorized into two groups depending on their onset times: congenital and adult-onset disorders; onset of the former is during infancy, whereas the latter manifest at older ages in adulthood. These diseases are attributed to point mutations in genes directly involved in either telomere regulation or DNA repair[174].

Dyskeratosis congenita is clinically typified by dermatological dystrophy and bone marrow deficiency, and commonly associated with aplastic anaemia and liver cirrhosis.

Missense variants that lead to partial loss-of-function of the dyskerin protein, resulting in impaired binding to telomerase RNA subunit, were found to be causal for X-inherited dyskeratosis congenita. Moreover, mutations that impair *TERC* gene transcription, leading to impaired enzymatic activity of telomerase, were found to cause 10% of autosomal dominant dyskeratosis congenita[173,174]. Hoyeraal–Hreidarsson syndrome is a severe form of the dyskeratosis congenita, showing the earliest disease onset time among all telomeropathies; *TIN2*, one of the structural component of SHELTERIN complex, was identified as the causal gene for this syndrome[174].

Ataxia telangiectasia, clinically presenting with hyper-radiosensitivity and immunodeficiency, is an autosomal recessive disorder. *Ataxia Telangiectasia Mutated* (*ATM)*, which encodes a serine/threonine protein kinase that phosphorylates and activates several key enzymes in DDR and cell cycle arrest, has been identified as the causal gene[173].

Fanconi anaemia, another autosomal recessive disorder, is caused by mutations in genes (including *BRCA2*) that encode proteins involved in protecting genome integrity from carcinogenic and oxidative stress. Pancytopenia and cancer susceptibility (acute myeloid leukaemia in particular) are two clinical phenotypes that manifest the disease, with significantly increased rates of cellular turnover and thus shorter TL[173,175].

WS and HGPS are the two best-characterised premature ageing syndromes, pathologically linked to drastic telomere attrition. Although their causal genes are neither direct regulators nor structural components of the telomeres, their genetic causes can indirectly affect telomere homeostasis. For example, the *WRN* gene, encoding the *RecQ* DNA helicase, is a major causal candidate for WS and likely to be involved in telomeric DNA replication; the *LMNA* gene, encoding a protein that contributes to formation of nuclear lamina matrix, has been identified as a major cause of HGPS, with functions in DNA repair and chromosomal stability, and thus is also essential for TL mainteinance[125].

Idiopathic pulmonary fibrosis, an autosomal dominant disease, with relatively higher prevalence in populations and a broader spectrum of ages of onset, can be ascribed to mutations in genes that encode telomerase enzymatic and RNA subunit, *TERT* and *TERC*, respectively[174].

1.4.1.4.2 Common complex diseases

TL has been consistently shown to be associated with various common complex diseases in epidemiological studies, including MR and observational studies[176].

MR studies have used summary statistics of genetic associations with LTL and various common complex diseases to estimate causal associations of LTL with different diseases[158,177–181]. This has suggested genetically predicted longer LTL is associated with increased risks of different types and sites of cancers (glioma, ovarian cancer, lung adenocarcinoma, neuroblastoma, bladder cancer, melanoma, testicular cancer, kidney cancer, and endometrial cancer)[158], and altered risks of several non-cancerous diseases, including increased risks of abdominal aortic aneurysm, celiac disease and interstitial lung disease[158] and decreased risks of coronary heart disease (CHD)[158,177–179] and AD and early-onset dementia[180,181].

In line with MR suggested protective effects of longer LTL on CHD and AD, recent meta-analyses of published observational studies have reported reverse associations of LTL with CHD[182] and AD[183]. However, contradictory results have been found on associations between LTL and cancers. For example, a large prospective cohort study has found longer LTL was associated with decreased risks of overall cancers and all-cause mortality during a 15-year follow-up[184,185]. Another study has shown that over 70% of cancer patients (n=9,127 patients, 31 cancers) exhibited shorter telomeres in their tumour samples in comparison to their matched normal tissue control samples, with the greatest differences found in melanoma, lymphoma and kidney tumours[186]. Moreover, patients with prostate cancers exhibited a U-shape of TL over the prostatic cancer development, with shorter telomeres during primitive or progressive stages, and longer telomeres during metastatic stages[187].

Based on previous studies, it is still unclear whether and how TL alteration is causally linked to different common complex diseases, including cancers and non-cancerous diseases. Observational studies can be subject to confounding and reverse causation, which are avoided in MR studies under the assumptions that genetic instruments are robustly and specifically associated with the exposure of interest. However, previous MR studies were performed based on summary results obtained from recent GWAS on LTL, which have only identified less than 10 genome-wide independent variants, explaining ~1% of total variation in LTL[151]. Larger GWAS meta-analyses can increase power of MR studies. Moreover, potential pleiotropic effects can be minimised through cross-referencing GWAS summary statistics for

thousands of traits for each variant. Overall, more robust and specific instrumental variables can lead to more precise causal estimates of LTL to various common complex diseases.

1.4.2 mLOY

1.4.2.1 Definition of mosaic chromosomal alterations and mLOY

Mosaic chromosomal alterations are post-zygotically acquired structural changes that can occur on any chromosomes, with a minimum size of 50bp[188,189]. It can arise early during development and thus influence both somatic and germline cells, or later in adulthood and influence only certain cell types[190]. It often originates during cell proliferation, where errors in chromosomal replication and subsequent transmission into daughter cells may lead to chromosomal anomalies that include copy number variation (CNV, segments or entire copies of chromosomes amplified or deleted) and uniparental disomy (two copies of an entire chromosome or sections of a chromosome coming from one parent)[189,190]. In principle, mosaic chromosomal alterations can lead to aberrant clonal expansions of progenitor and stem cells carrying structural alterations that provide cellular growth advantages[191]. The aberrant clonal expansions can occur in all tissues regardless of the clonal or developmental origin of cell lineages[189]. In particular, clonal expansions among hematopoietic stem cells are referred to as the "clonal hematopoiesis", which is the most extensively studied type of mosaicism, possibly due to an easier accessibility to blood DNA samples[192], although no significant differences were found between DNA extracted from different samples, such as blood and buccal cells[189]. mLOY is a particular type of mosaic chromosomal alterations that occurs on the chromosome Y. mLOY is by far the most common mosaic chromosomal alteration present in men[193].

1.4.2.2 Germline genetic variants associated with mosaic chromosomal alterations and mLOY

Mosaic chromosomal alterations by definition were non-heritable due to somatic nature of the mutations that drive disproportionate expansions of clonal lineages; however, several large-scale GWAS have demonstrated that inherited germline variants can influence risk and chromosomal positions of such mosaic loss[194–196].

While most of the previous genetic studies on mosaic chromosomal alterations were restricted to only a limited number of mosaic events that have been demonstrated to be potentially pathogenic, several studies have used a systematic approach and analysed mosaic chromosomal alterations at a genome-wide scale[194,197–199]. The clonal mosaic events analysed in these studies were either aneuploidy or copy-neutral loss of heterogeneity, with scales ranging from 50kb to a whole chromosome. For *cis*-associations, six germline variants have been linked to nearby mosaic chromosomal alterations, including those near *FRA10B*, *MPL*, *ATM*, *TM2D3/TARSL2*, *DXZ1* and *DXZ4* genes[194]. For *trans*-associations, a variant at the *SP140L* gene and a variant within the *HLA* region have been associated with mosaic loss of chromosome X (mLOX)[194]. Moreover, a more specific search within a subset of carcinogenic variants identified *TERT*, *TP53* and *CHEK2* genes to be associated with multiple types of clonal mosaic events[194].

Several recent studies have performed GWAS on mLOY in leukocytes at increasingly larger scales. The first GWAS on this trait found an common variant located at the 5' end of *TCL1A* gene, an oncogene that causes T cell leukemia[196]. More recently, a GWAS meta-analysis combining samples from UK Biobank, EPIC (European Prospective Investigation of Cancer)-Norfolk and deCODE (Diabetes Epidemiology: Collaborative analysis of Diagnostic criteria in Europe) cohorts has largely expanded the number of genomic regions associated with mLOY to a total of 19, explaining 2.7% of total variance in mLOY[195]. Functional analyses of candidate genes within these regions highlighted pathways involved in DNA repair and cell cycle regulation[195]. Applying a newly developed algorithm of mLOY detection within UK Biobank, a study has identified 156 autosomal genetic determinants of mLOY, highlighting novel candidate genes functioning in cell-cycle regulation and cancer susceptibility[135].

1.4.2.3 Non-genetic risk factors associated with mosaic chromosomal alterations and mLOY

Mosaic chromosomal alterations occur in an age-dependent manner in normal healthy populations[200]. Structural *de novo* variants (>50kb) occur more frequently in elderly populations, and the frequencies increase proportionately to age between 50 and 80 years[201]. Other studies have reported consistent findings showing that the frequencies of post-zygotic structural variants increase along with age in healthy individuals[189,190,194,202–206]. Moreover, distributions of clonal mosaic events were non-random, which exhibited region-specific stratifications by age and sex, and higher frequencies around carcinogenic genes[194].

Other haematological, behavioural phenotypes and clinical treatments have also been associated with clonal mosaicism, including multiple blood cell components, smoking and therapeutic treatments for addiction and psychiatric disorders[206].

mLOY exhibited similar distributions and phenotypic associations to the overall mosaic chromosomal alterations[206]. Several studies have demonstrated that older age and current smoking status were strong risk factors for mLOY, and suggested a causal effect of current smoking on mLOY[193,195,196].

1.4.2.5 Clinical consequences of mosaic chromosomal alterations and mLOY

A variety of clinical consequences can be attributed to mosaic chromosomal alterations and mLOY in particular, including cancers, and age-related diseases, such as AD and cardio-metabolic disorders[189,190,207].

Mosaic chromosomal alterations often affect specific genes that are involved in cancer development[200]. For instance, somatic mutations in *DNMT3A*, *ASXL1*, and *TET2* genes have frequently been observed in detectable clonal expansions, which have previously been implicated in haematological malignancies[192]. Aberrant clonal expansions of hematopoietic cells were strongly linked to carcinogenesis, and often acknowledged as pre-cancerous states, with evidence supported by a clinical study that showed ~42% of haematological cancer patients (12,380 individuals from the Swedish national patient registers without prior selection for haematological phenotypes) exhibited clonality more than 6-months before their first cancer diagnoses[192]. Clonal hematopoiesis has been shown to be associated with more than 10-fold higher risks of cancers in haematological tissues[194], as well as cancers in several non-haematological tissues, yet with smaller effect sizes[189,190,193,205]. Similar to the overall mosaic alterations, mLOY has been associated with cancer-related and all-cause mortality[193], and risks of cancers at specific sites, such as bladder and prostate but not lung tissues[196].

Non-cancer related diseases have also been reportedly associated with mosaic chromosomal alterations and mLOY, such as CVD and AD. Clonal haematopoiesis has been suggested to increase risks of CHD and ischemic stroke by approximately 2-fold, as well as early-onset myocardial infarction by 4-fold on average[17,19]. mLOY has been associated with increased risk of late onset, sporadic AD[207], implying a role of mLOY in age-related neurodegenerative disorders. Compared to the implication of mosaic events in cancers, their

relevance to non-cancerous age-related diseases, such as CVD, AD and T2D has been less studied. How mosaic chromosomal alterations, and mLOY in particular, influence risks of these age-related common diseases may be better elucidated in large prospective cohorts.

Chapter 2

GWAS of LTL

Abstract

Background LTL is a highly heritable trait, yet previous GWAS have identified only a small number of genetic loci, with a large proportion of the heritability unexplained.

Objectives To expand our current knowledge of genetic regulation of LTL, and propose likely causal mechanisms underlying such regulation.

Methods Genome-wide association analyses were performed in EPIC-InterAct, EPIC-CVD and ENGAGE studies followed by meta-analyses, accumulating to 78,592 individuals with densely imputed genotypes. Causal gene candidates were prioritised via multi-omic data integration using methods, such as S-PrediXcan and statistical co-localisation. Pathway enrichment analyses were performed to identify potential biological mechanisms that underpin genetic regulation of LTL.

Results There were 17 genomic regions associated with LTL at genome-wide significance level, among which 6 were novel, located in or near *SENP7*, *MOB1B*, *CARMIL1*, *PRRC2A*, *TERF2, RFWD3* genes. Moreover, there were 32 additional regions identified at FDR<0.05. In total, we prioritised 32 causal gene candidates, which were functionally enriched in pathways involving telomere structure and maintenance, DNA damage response, and nucleotide metabolism.

Conclusions Our findings increase the total number of genomic regions associated with LTL, with a more comprehensive elucidation of the genetic architecture, and provide better characterisation of likely causal genes and biological mechanisms underlying regulation of telomere homeostasis.

2.1 Introduction

Telomeres are ribonucleoprotein complex located at the end of chromosomes, regulating cell division and genome integrity, as outlined in more detail above (section 1.4.1.1). TL, most commonly measured in human leukocytes (LTL), displays large variation between individuals, from birth and throughout the life course, yet is highly heritable, with heritability estimates between 44-86%[128,171]. GWAS so far have Identified 10 genetic regions associated with LTL, with a large proportion of the heritability unexplained. Identification of genetic determinants of LTL through GWAS has enabled further studies to suggest a causal role of LTL in several diseases, including CAD, abdominal aortic aneurysm, various cancers, interstitial lung disease and celiac disease[151,179,180,208,209]. These studies are however limited due to the small number of genetic variants that have been identified that replicate between studies[96,151–154,157,210,211]. To further our understanding of LTL regulation and its relationship with diseases we conducted a GWAS meta-analysis of 78,592 individuals from ENGAGE (European Network for Genetic and Genomic Epidemiology), EPIC-CVD and EPIC-InterAct studies. This chapter focuses on the genetic discovery part of the GWAS, including identification of genetic variants, and their functional implications with telomere biology via causal gene characterisation and pathway enrichment analyses, while clinical relevance of LTL will be covered in the next chapter.

2.2 Methods

2.2.1 Study design

We used data from EPIC-InterAct, EPIC-CVD and ENGAGE studies to perform GWAS on standardised mean LTL in up to 79,000 individuals (Supplementary Figure 1). Detailed description and demographic characteristics of all study cohorts, for both discovery and replication phases are shown in Supplementary Notes and Supplementary Table 1. In brief, both EPIC-InterAct and EPIC-CVD are case-cohort studies, with focuses on incident T2D and CVD respectively. EPIC-InterAct consists of 12,403 ascertained cases of T2D and a quasi-

random sub-cohort of 16,154 participants[212,213]. EPIC-CVD uses the same sub-cohort as InterAct, and thus participants included in this analysis are incident cases only (7722 CHD cases and 3451 cerebrovascular disease cases)[214]. ENGAGE study consists of 21 independent cohort studies across European countries, which has been previously described[151], and 3 additional studies included in this meta-analysis are GENMETS[215], a Finnish population-based cohort study of T2D cases and controls; NESDA (the Netherlands Study of Depression and Anxiety)[216] and Rotterdam Study that investigates occurrence and determinants of diseases in the elderly[217]. All individuals included in the analyses are of European descent and provided written informed consent.

GWAS were performed separately for each study or subset/stratum contributing to either EPIC-InterAct, EPIC-CVD or ENGAGE, followed by inverse variance weighted meta-analyses, as outlined in more detail below (section 2.2.3). Compared to the previous publication by the ENGAGE consortium, this GWAS meta-analysis more than doubled the previous sample size, and largely expanded SNP coverage by ~5-fold via upgrading the imputation reference panel from HapMap II to HRC and 1000G, increasing the total number of genetic variants to over 10 million (section 2.2.2)[151]. A systematic conditional analysis was performed for each locus at FDR<0.05 (section 2.2.5), followed by cross-platform bioinformatic annotations of independent variants at each locus as well as their closely correlated variants (LD r^2>0.8). Causal gene candidates were prioritised using a variety of computational prediction methods that integrate transcriptional and epigenetic data and validated by knowledge-driven manual curation. Pathway enrichment analyses were performed to characterise functional commonalities shared between prioritised genes. Clinical consequences of genetic variations in LTL were tested in a hypothesis-free, phenome-wide scan in UK Biobank (chapter 3).

2.2.2 LTL Measurements and QC analysis

LTL measurements were conducted using an established quantitative polymerase chain reaction (PCR) technique which expressed TL as a ratio of the telomere repeat numbers (*T*) to the single copy of a housekeeping gene (*S*)[42,43]. LTL measurements were standardised using either a calibrator sample or by quantifying against a standard curve, depending on laboratories (Supplementary Table 1 and Supplementary Notes). Full details of methods

employed by different laboratories, along with QC parameters, are given in the Supplementary Notes. As the use of different calibrator samples and standard curves can lead to different ranges in *T/S* ratios observed between laboratories, we standardised LTL using a z-transformation approach ($z = (\mu - \mu_0)/\sigma$, μ, *T/S* ratio, μ_0, mean of *T/S* ratio, σ, standard deviation (SD)).

2.2.3 Genotyping, GWAS analysis and study level QC

Genotyping platforms and imputation methods and panels varied across participating study centres. Detailed information about these is provided in Supplementary Figure 1 and Supplementary Table 2. GWAS was performed within each contributing study or subset/stratum: for EPIC-InterAct and EPIC-CVD studies, analyses were stratified by disease status (incident T2D cases, incident CVD cases, control cohort participants) and genotyping platforms (Human CoreExome, Illumina-660W-Quad and HumanOmniExpress), resulting in 9 individual GWAS for the EPIC meta-analysis; for ENGAGE consortium cohorts, a total of 24 contributing studies were analysed separately and then meta-analysed. We used linear regression under an additive mode of inheritance with adjustment for age, sex and study specific covariates including batch of LTL measurement, study centre and genetic PCs. Within each study or subset/stratum, related samples ($k>0.088$) were removed. Population stratification was estimated using the genomic control inflation factor, λ (QQ plots were shown in Supplementary Figure 2), and used to adjust standard errors of results from each GWAS. Genetic variants were kept based on standard criteria including call rates > 95%, Hardy–Weinberg equilibrium *p*-value > 1×10^{-6}, imputation quality scores > 0.4 or $R^2>0.3$, minor allele counts $\geqq$ 10 and standard errors of association estimates ranging from 0 to 10[151,213,214]. Results meeting these criteria were taken forward to meta-analyses.

2.2.4 Meta-analyses

GWAS summary statistics were combined via two steps of meta-analyses using inverse variance weighting in GWAMA (Genome-Wide Association Meta-Analysis)[218]. In the first step, we performed the EPIC meta-analysis across 9 subsets/strata that contribute to the EPIC-

InterAct and EPIC-CVD studies, and a separate meta-analysis across all 24 cohort studies within the ENGAGE consortium. Fixed effects were used except for variants with significant heterogeneity (Cochrane's Q: *p*-value<$1x10^{-6}$) where random effects were used. Genomic control was applied at this step of meta-analysis. Genetic variants that had more than 40% of the total sample size in each of these two meta-analyses were retained. In the second step, we meta-analysed results from EPIC meta-analysis with those from ENGAGE using fixed effects inverse variance weighted method. No genomic control was applied at this step. We calculated FDR by estimating q-values[219].

2.2.5 Conditional association analysis

Conditionally independent signals were identified by an approximate genome-wide stepwise method using GCTA (Genome-wide Complex Trait Analysis, Version 1.25.2)[220,221]. Summary statistics for SNPs included in the final step of meta-analysis were used as the input, with a *p*-value cut-off of $1.03X10^{-5}$ (FDR=0.05) used to indicate regional significance level. The model starts with the most significant SNP, with more SNPs added iteratively in a forward stepwise manner, and conditional *p*-values calculated for all SNPs considered within the model. This forward selection process was repeated until no more SNPs can be added into the model, i.e. no more added SNPs that can reach the conditional *p*-value threshold. During the selection process, if SNPs show evidence of collinearity (LD r^2>0.9, estimated based on a random subcohort of UK biobank, n=50k) with any of the existing SNPs in the model, those SNPs will be automatically dropped and excluded from the model[220]. Joint effect of each selected SNP in the model was calculated and reported as conditionally independent effect of that SNP. Regional plots of genome-wide significant loci were generated using LocusZoom[222] with LD structure estimated using the random subcohort of UK biobank (Supplementary Figure 3).

2.2.6 Gene prioritization

2.2.6.1 Variant annotation

Variants (conditional *p*-value < 1.03 X 10^{-5}) and their closely related variants (LD r^2>0.8) were annotated on the human reference genome sequence hg19 using Annovar (v2017July16)[223] and Variant Effect Predictor (VEP)[54]. Their functional consequences on protein sequences encoded by the nearest genes were cross-referenced by definitions from RefGene[224], Ensembl gene annotation[225], GENCODE[226] and UCSC human genome database[55]. These variants were also evaluated for features (UCSC genome database) including evolutionary conservation: whether they reside in or specifically encode a conserved element by multiple alignments across 46 vertebrate species, chromatin states predicted using Hidden Markov Models trained by CHIP-seq (CHromatin ImmunoPrecipitation assays with sequencing) data from ENCODE (ENCyclopedia Of DNA Elements, 15 classified states across 9 cell types), histone modification markers (active promoter: H3K4Me3, H3K9Ac; active enhancer: H3K4me1, H3K27Ac; active elongation: H3K36me3; and repressed promoters and broad regions: H3K27me3) and CCCTC-binding Factor (CTCF) transcription factor binding sites across 9 cell lines, conserved putative transcription factor binding sites (TFBS) and DNaseI hypersensitive areas curated from ENCODE database. For variants within exons, they were further annotated with allele frequencies from 7 ethnic groups from the Exome Aggregation Consortium database, and functional predictions using a number of different algorithms (Supplementary Table 6). For non-coding variants we performed integrated analysis with SNP Nexus IW scoring[227] (Supplementary Table 8).

2.2.6.2 Transcriptomic data integration

(1) With summary statistics, we performed a gene-level analysis using S-PrediXcan that links LTL to predicted gene expressions across 44 tissues (GTex v6p). It uses multivariable sparse regression models that integrate *cis*-SNPs within 2Mb windows around boundaries of gene transcripts to predict corresponding gene expression levels. Detailed description of the method can be found elsewhere[60,228]. In brief, individual SNP-LTL associations were weighted by SNP-gene (w_{lg}) and SNP-SNP ($\frac{\sigma l}{\sigma g}$) association matrices estimated from the PredictDB training set ($z_g = \sum_{l \in g} w_{lg} \frac{\sigma l}{\sigma g} z_l$, for a gene ($g$); the set of SNPs ($l$) were selected from an elastic net model with a mixing parameter of 0.5). Protein-coding genes with qualified prediction model performance (average Pearson's correlation coefficients r^2 between predicted and

observed gene expression levels >0.01, FDR<0.05) were included in our analysis. We considered a predicted gene expression to be significantly associated with LTL at a Bonferroni corrected *p*-value threshold (*p*-value<2.61 X 10^{-7}), conservatively assuming association test for each gene-tissue pair as an independent test.

(2) Because the S-PrediXcan analysis may contain false positive findings especially when LD structures do not match closely between populations where the SNP-LTL and the SNP-gene association matrices were estimated, and/or tissue sample sizes are small[228], we used another eQTL integration method to calibrate S-PrediXcan results, the statistical colocalisation method[63]. This method, using a COLOC Bayesian approach, tested whether there was evidence for potential causal variants being shared between LTL and gene expression levels. We analysed all loci significantly associated with LTL at FDR<0.05, and defined a region for testing as a 2Mb window flanking the lead variant of that region. Regional summary statistics were extracted from this GWAS meta-analysis for associations with LTL and GTex v7[59] for *cis*-eGenes (genes with at least one significant eQTL at FDR < 0.05) located within or on boundaries of LTL loci identified at FDR < 0.05. Default priors were applied (prior probabilities of SNPs to be associated with either LTL or gene expressions to be 1X10^{-4} and with both to be 1X10^{-5}). Evidence for colocalisation was assessed by comparing PPAs for two hypotheses: associations with both traits were driven by the same causal variant (hypothesis 4) and by distinct variants (hypothesis 3). We defined an eGene as having evidence of genetic colocalisation with LTL when the ratio of PPA for the hypothesis 4 to the sum of PPAs for both hypotheses 3 and 4 was larger than 0.9.

2.2.6.3 Epigenomic (DNA methylation) data integration

For genes whose expressions are modulated by epigenetic modifications, such as methylation of transcriptional regulators in *cis*, integrating genetic associations with *cis*-methylation probes (*cis*-meQTLs at FDR<0.05) can help prioritise causal gene candidates with evidence from epigenetic transcription modulation. For this, I conducted 1) a systematic search of lead variants of LTL-associated loci and their proxies (r^2>0.8) in multiple publicly available meQTL databases[229–231]. 2) an epigenome-wide association scan that integrates multiple variant associations with *cis* methylation probes with those with LTL, using a regularised linear

regression model, which algorithmically is similar to a transcriptome-wide association analysis, previously described in detail elsewhere[61]. A reference panel for the *cis*-meQTLs was constructed based on individuals in the EPIC-Norfolk cohort, with detailed descriptions published elsewhere[195]. Bonferroni correction was applied, accounting for the total number of CpG markers tested (*p*-value=1.00×10^{-7}).

2.2.8 Pathway enrichment analysis

2.2.8.1 Protein ANalysis THrough Evolutionary Relationships (PANTHER)

A list of prioritised genes at each locus (or the nearest gene where no prioritization was possible) was submitted for statistical overrepresentation testing (Fishers exact test) in PANTHER[232]. Pathways were considered over-represented where FDR<0.05.

2.2.8.2 Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)

I also used a hypothesis-free, data-driven approach to highlight reconstituted gene sets and tissue/cell types where LTL-associated loci were enriched using DEPICT. Detailed description of gene set construction has been published elsewhere[233]. Briefly, DEPICT leveraged a broad range of pre-defined pathway-oriented databases to construct gene sets (14,461), including GO (Gene Ontology) terms[234], KEGG (Kyoto Encyclopaedia of Genes and Genomes) database[235], Reactome pathways[236], experimentally-derived protein-protein interaction sub-network[237] and a gene-phenotype matrix curated by Mouse Genetics Initiative[238]. Summary statistics of uncorrelated SNPs (LD $r^2 \leqq 0.5$) significantly associated with LTL at a genome-wide level (*p*-value<$5X10^{-8}$) were used as input, with the HLA region (chr6:29691116–33054976) excluded. DEPICT first characterised gene functions based on pairwise co-regulation of gene expressions, which were quantified as membership probabilities across all reconstituted gene sets. Then for each gene set, it assessed enrichment by testing if the sum of membership probabilities of all genes within each LTL-associated locus was higher than that for a gene density-matched random locus. Correlations (Pearson's $r^2 \geqq 0.3$) between significant gene sets were visualised using CytoScape[239].

2.3 Results

2.3.1 Discovery of novel genetic determinants of LTL

In total, 20 sentinel variants at 17 genomic loci were independently associated with LTL at a level of genome-wide statistical significance (*p*-value<$5x10^{-8}$, Table 2.1, Supplementary Figure 2), including 6 novel loci [*Sentrin/SUMO Specific Peptidase 7* (*SENP7*), *MOB Kinase Activator 1B* (*MOB1B*), *Capping Protein Regulator And Myosin 1 Linker 1* (*CARMIL1*), *Proline Rich Coiled-Coil 2A* (*PRRC2A*), *TERF2, Ring Finger And WD Repeat Domain 3* (*RFWD3*)]. We also identified genome-wide significant variants in 4 recently reported loci from a Singaporean Chinese population [*POT1*, *Poly(ADP-Ribose) Polymerase 1* (*PARP1*), *ATM* and *M-Phase Phosphoprotein 6* (*MPHOSPH6*)][240] (Supplementary Table 5) and confirmed associations at 7 previously reported loci in European ancestry studies (*TERC*, *NAF1*, *TERT*, *OBFC1*, *ZNF208*, RTEL1, and *DDB1 And CUL4 Associated Factor 4* (*DCAF4*)*)*[151,153]. Two and three conditionally independent signals were detected within the *TERT* and *RTEL1* loci respectively (section 2.2.5, Table 2.1). Within the known loci, three variants within the *DCAF4* (r^2=0.05) and *TERT* (r^2<0.5) loci were distinct from previously reported sentinel variants, while five (r^2>0.8, *TERC*, *NAF1*, *OBFC1*, *ZNF208* and *RTEL1*, Supplementary Table 3) were in high LD with the previously reported ones from European studies. For loci identified in the Singaporean Chinese population, we observed the same sentinel variant for *PARP1*, high LD variants for *ATM* and *MPHOSPH6* (r^2>0.8) but a distinct sentinel for *POT1* (r^2<0.5, Supplementary Table 5). For the *RTEL1* locus, there were significant differences in LD structures between ancestral populations. All variants within the *RTEL1* locus we reported at genome-wide statistical significance were in low LD with those reported in the Singaporean Chinese study[211,240]. Our novel variants were of lower frequency (MAF<0.1) and were either reported as being monoallelic (monomorphic) or fell below the MAF threshold for analysis in the Singaporean Chinese population (MAF<0.01). This suggested that genetic variation in this region may be, in part, population specific or that the MAF was so low that we currently were unable to detect any associations.

A total of 32 additional variants met the FDR threshold of 0.05 (Supplementary Table 4). These variants were located within separate loci to those reported above, with the exception of a fourth, independent signal in the *RTEL1* locus. Although we did not replicate the previously reported *ACYP2* locus at genome-wide significance, this remained within variants identified at the FDR<0.05 threshold. *Thymidylate Synthetase* (*TYMS*), identified as genome-wide significant in a trans-ethnic meta-analysis of the Singaporean Chinese[240] and the previously reported ENGAGE analysis[151], was also within the FDR<0.05 identified loci. This was to be expected considering a substantial sample overlap with the ENGAGE data, however our sentinel variant was distinct and not reported in the Singaporean Chinese study. Aligning our data to available summary statistics from that study (Singaporean Chinese samples only) we saw at least nominal support for the vast majority of our genome-wide significant loci, with the exception of *OBFC1* and *SENP7* (Supplementary Table 5). Whilst *SENP7* is a novel locus, variants in high LD (r^2>0.6) with our *OBFC1* sentinel variant have been reported in other European populations[154,210]. There is also support for many variants in our extended FDR list. However, it should be noted that data was not available for around half of our FDR<0.05 loci, with most of these being either monoallelic or of too low frequency to have been included within the analysis in the Southern Chinese Han population, again suggesting several of the FDR<0.05 loci may be specific to the European population.

Table 2.1 Independent variants associated with LTL at genome-wide significance (*p*-value=5X10^{-8}).

Additional, independent signals detected using conditional analysis are included*. Gene denotes the closest or candidate genes within the region. EA is effect allele, EAF is effect allele frequency within the study, Beta is per-allele effect on z-scored LTL and SE is standard error. Previously reported loci were defined as loci reported before 1st September 2019.

	SNP	Gene	Chr	Position	EA	EAF	Beta	SE	*p*-value
Previously reported loci	rs3219104	PARP1	1	226562621	C	0.83	0.042	0.006	9.60E-11
	rs10936600	TERC	3	169514585	T	0.24	-0.086	0.006	7.18E-51
	rs4691895	NAF1	4	164048199	C	0.78	0.058	0.006	1.58E-21
	rs7705526	TERT	5	1285974	A	0.33	0.082	0.006	5.34E-45
	rs2853677*	TERT	5	1287194	A	0.59	-0.064	0.006	3.35E-31
	rs59294613	POT1	7	124554267	A	0.29	-0.041	0.006	1.17E-13
	rs9419958	STN1 (OBFC1)	10	105675946	C	0.86	-0.064	0.007	5.05E-19
	rs228595	ATM	11	108105593	A	0.42	-0.029	0.005	1.43E-08
	rs2302588	DCAF4	14	73404752	C	0.10	0.048	0.008	1.68E-08
	rs7194734	MPHOSPH6	16	82199980	T	0.78	-0.037	0.006	6.94E-10
	rs8105767	ZNF208	19	22215441	G	0.30	0.039	0.005	5.42E-13
	rs75691080	RTEL1/STMN3	20	62269750	T	0.09	-0.067	0.009	5.99e-14
	rs34978822*	RTEL1	20	62291599	G	0.02	-0.140	0.023	7.26E-10
	rs73624724*	RTEL1/ZBTB46	20	62436398	C	0.13	0.051	0.007	6.33E-12
Novel loci	rs55749605	SENP7	3	101232093	A	0.58	-0.037	0.007	2.45E-08
	rs13137667	MOB1B	4	71774347	C	0.96	0.077	0.014	2.43E-08
	rs34991172	CARMIL1	6	25480328	G	0.07	-0.061	0.011	6.19E-09
	rs2736176	PRRC2A	6	31587561	C	0.31	0.035	0.006	3.53E-10
	rs3785074	TERF2	16	69406986	G	0.26	0.035	0.006	4.64E-10
	rs62053580	RFWD3	16	74680074	G	0.17	-0.039	0.007	4.08E-08

2.3.2 Prioritization of likely causal genes

We applied *in silico* prediction tools, leveraging large-scale human genomic data integrated with multi-tissue gene expression, transcriptional regulation and DNA methylation data, and knowledge-driven manual selection to prioritise likely-causal genes (section 2.2.6, Supplementary Tables 7, 9 and 10, and Supplementary Notes). Where prioritisation methods suggested multiple causal genes for a given locus, we pinpointed the most probable one, where feasible, as that showing the greatest number of lines of evidence with supports from manual annotation (Supplementary Notes). We were able to predict likely causal genes at 15 of the 17 loci at genome-wide significance and 17 of the 32 loci at FDR<0.05, many of which were for the first time linked to telomere biology, providing novel insights into gene functions that are potentially implicated in TL regulation (Table 2.1, Supplementary Table 10).

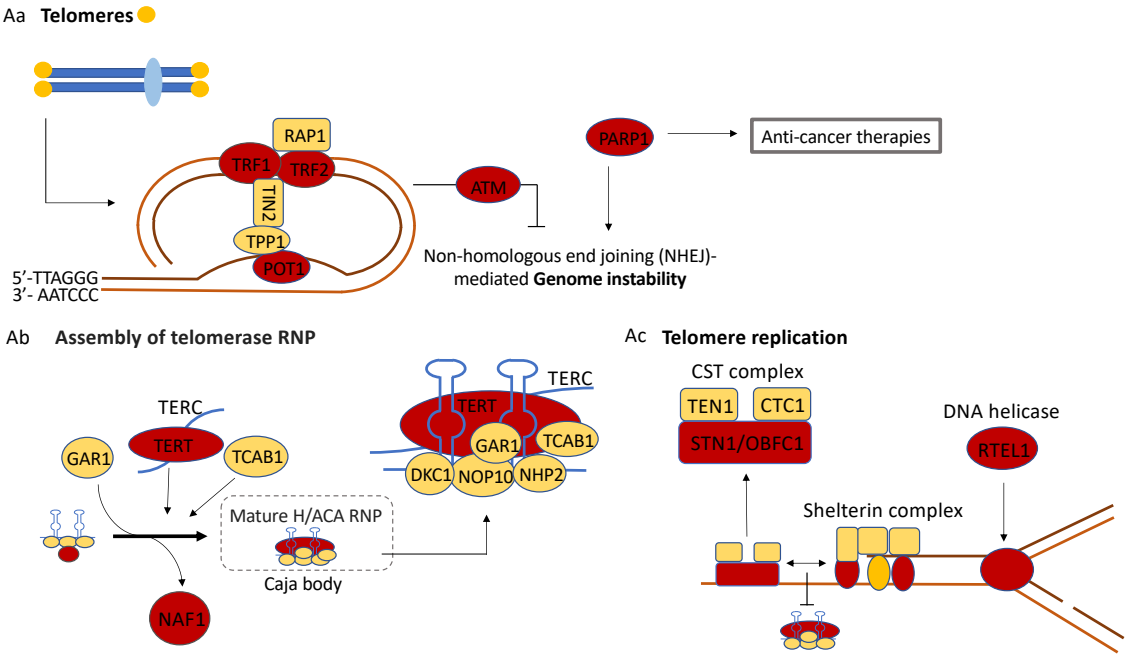
Within the novel loci, four genes have known roles in TL regulation (*PARP1*, *POT1*, *ATM*, *TERF2*, Figure 2.1). For *PARP1*, a variant in complete LD with our identified sentinel variant causes the V762A substitution (Supplementary Table 6) known to reduce PARP1 activity[34]. This variant was associated with shorter LTL, in agreement with studies showing that knock down of *PARP1* led to telomere shortening[35,36].

Three genes, *DCAF4*[30-31], *SENP7*[31] and *RFWD3*, prioritised based on deleterious protein coding changes (*DCAF4*, *SENP7*) or strong evidence linking to gene expression levels (*RFWD3*) were all involved in DNA repair. SENP7 has been demonstrated to have binding affinity to damaged telomeres[33], giving further credibility to this gene. Components of DDR and repair pathways (such as ATM) have previously been shown to play roles in telomere regulation[32].

The *PRRC2A* locus contains 11 genetically-linked SNPs located across the major histocompatibility complex (MHC) class III region, which is a highly polymorphic and gene-dense region with complex LD structure. *BAG6* and *CSNK2B* were suggested as causal gene candidates for this region, which were supported by evidence showing associations of their genetically predicted gene expressions with LTL (Supplementary Notes, Table 2.1, Supplementary Table 7). BAG6 is linked to DNA damage signalling and apoptosis[241], whilst CSNK2B, a subunit of casein kinase 2, binds to TERF1 and regulates telomere homeostasis[242].

Figure 2.1. Loci with established roles in telomere biology.

Candidate genes found in this study are in red, genes not identified in this GWAS meta-analysis are in yellow. Candidate genes include genes that encode components of the SHELTERIN complex (Aa), regulate the formation and activity of telomerase (Ab), and telomere replication (Ac).



2.3.3 Pathway enrichment

To investigate context-specific functional connections between prioritised genes of the 17 genome-wide significant loci and suggest plausible biological roles of these genes in TL regulation, we performed enrichment analyses for pathways using DEPICT and PANTHER (Section 2.2.8). Over 300 reconstituted gene sets (DEPICT) were significantly (FDR<0.05) enriched for the LTL loci, which can be further clustered into 34 meta-gene sets, highlighting pathways that are involved in several major cellular activities, including DNA replication, transcription and repair, cell cycle regulation, immune response and intracellular trafficking (Figure 2.2A).

The PANTHER analysis identified several telomere related pathways, including regulation of telomeric loop disassembly, t-circle formation, protein binding at telomeres and single strand break repair as being the most highly overrepresented. Amongst other expected pathways, cellular ageing and senescence were also highlighted. Of note, nucleotide metabolism pathways were overrepresented (2'-deoxyribonucleotide metabolic process, deoxyribose phosphate metabolic process, deoxyribonucleotide metabolic process, Figure 2.2B, Supplementary Table 11). Three genes were matched to this pathway, *SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1* (*SAMHD1*), *single-strand selective monofunctional uracil DNA glycosylase 1* (*SMUG1*) and *TYMS*. In addition, two further genes within other identified loci, *deoxycytidine kinase* (*DCK*), *thymidine kinase 1* (*TK1*), were key regulators of dNMP biosynthesis, even though not highlighted in the pathway analysis, adding further support to nucleotide metabolism as a key pathway in regulating LTL[243]. dNTPs constitute the fundamental building blocks required for DNA replication and repair[244,245]. Genetic perturbations that disrupt dNTP homeostasis have been shown to result in increased replication error, cell cycle arrest and DNA damage induced apoptosis[246–248].

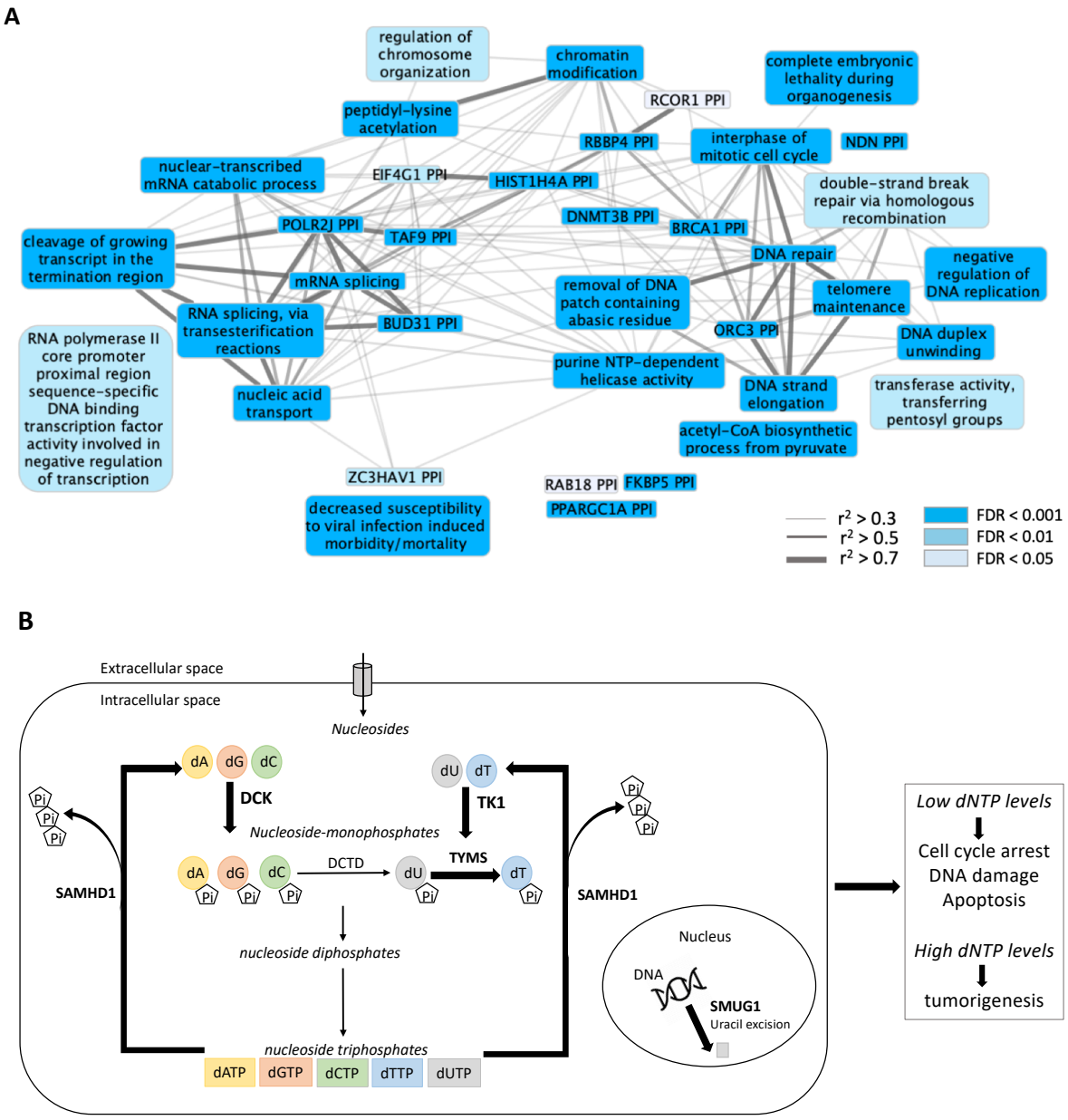
TK1 and DCK are the rate-limiting enzymes that catalyse the first step of the salvage pathway of nucleotide biosynthesis, either phosphorylating deoxythymidine (dT) to produce dTMP, or deoxycytidine (dC), deoxyguanosine (dG) and deoxyadenosine (dA) to dCMP, dGMP and dAMP respectively (Figure 2.2B). The salvage pathway relies on extracellular nucleosides originated from diet or dephosphorylation of recycled nucleotides, whereas the *de novo* pathway utilises basic constituents, including glucose and amino acid derivatives[249]. TYMS is considered as a component of the *de novo* pathway, converting dUMP to dTMP, where the

dUMP substrates can be derived either from *de novo* synthesis or deamination of dCMP (Figure 2.2B).

Besides controlling biosynthetic pathways, equilibrium of cellular dNTP levels is also achieved by regulating degradation that prevents overproduction of dNTPs, where an enzyme encoded by another prioritised gene, *SAMHD1* plays a role. It catalyses hydrolysis of dNTPs to deoxynucleosides and triphosphates, thereby preventing accumulation of excess dNTPs[250] (Figure 2.2B). The finely-tuned dNTP supply system inhibits incorrect insertion of bases into DNA synthesis, which is also monitored by a base excision repair enzyme, *SMUG1* that removes uracil and oxidised derivatives from DNA molecules[248,251] (Figure 2.2B).

Figure 2.2 Pathways Enriched for Telomere-associated Genes.

A. Gene sets significantly (FDR<0.05) enriched for prioritised LTL-associated genes. Colour intensity of nodes (gene sets) was classified into three levels, reflecting enrichment strengths (FDR). Edge width indicates Pearson correlation coefficient (r^2) between each pair of the gene sets. **B.** Role of telomere-associated genes in nucleotide metabolism. Five enzymatic reactions and corresponding enzymes encoded by genes prioritised from this GWAS are highlighted in bold.



2.4 Discussion

In this chapter, I present genome-wide genetic association findings for LTL, biological relevance of prioritised genes and their functional interconnections. The clinical consequences of TL dysregulation will be elucidated in detail in the subsequent chapter.

2.4.1 Discovery of novel variants within loci containing known telomere-related genes

Using this larger, and more densely imputed dataset we identified 20 lead variants at level of genome-wide significance and a further 32 at FDR<0.05. Within established loci we reported a second, independent, association signal within the *TERT* locus and redefined the *RTEL1* locus into three independent signals. By applying a range of *in silico* tools that integrate multiple lines of evidence, we were able to pinpoint likely causal genes for the majority of independent lead variants (32 of 52), several of which represent key telomere regulating pathways (including components of the telomerase complex, the telomere binding SHELTERIN and CST complexes and the DDR pathway).

Telomeres function to prevent 3' single-stranded overhangs at the end of chromosomes from being detected as double-stranded DNA breaks. This is achieved through binding of the SHELTERIN complex (TERF1, TERF2, RAP1, TIN2, TPP1 and POT1) which acts to block activation of DDR pathways by several mechanisms[252]. SHELTERIN also binds a number of accessory factors that facilitate processing and replication of the telomere, including the DNA helicase RTEL1[253]. SHELTERIN also interacts with the CST complex that regulates telomerase access to telomeric DNA[254]. The identified loci contain two of the SHELTERIN components (TERF2, POT1), a regulator of TERF1, CSNK2B at the *PRRC2A* locus[255], the helicase RTEL1 and the CST component STN1.

Whilst telomere binding proteins and structure aims to inhibit activation of DDR pathways, there is also evidence of a paradoxical involvement of a number of DDR factors in TL maintenance, including both of the prioritised genes *ATM* and *PARP1*[256–259]. TERF2 inhibits ATM activation and classical non-homologous end joining (c-NHEJ) at telomeres, and thus preventing synapsis of chromosome ends[260–262]. However, ATM activation is required for telomere elongation, potentially by regulating access of telomerase to telomeres end through ATM-mediated phosphorylation of TERF1[256,257]. It is possible that other DDR regulators can

impact TL maintenance through regulating telomeric chromatin states, t-loop dynamics and single-stranded telomere overhang processing[263]. Other prioritised genes (*SENP7*, *RFWD3*) also function within the DDR pathways, suggesting a plausible mechanism through which they may influence LTL.

The telomerase enzyme is capable of extending telomeres and/or compensating sequence loss due to the end replication problem in stem and reproductive cells[264,265]. Associated loci include genes encoding the core telomerase components TERT and TERC along with the chaperone protein NAF1. NAF1 is required for TERC accumulation and its incorporation into the telomerase complex[266]. After transcription, TERC undergoes complex 3' processing to produce the mature 451bp template[267,268]. This post-transcriptional process involves components of the RNA exosome complex, PARN (Poly(A)-Specific Ribonuclease) and PAPD5 (PAP-Associated Domain-Containing Protein 5, also known as the Non-Canonical Poly(A) RNA Polymerase PAPD5) amongst others; this process is not fully understood[269–271]. In addition to variants within regions containing *TERT*, *TERC* and *NAF1*, a prioritised gene from another locus (*MPHOSPH6*) is a component of the RNA exosome[272].

2.4.2 Nucleotide metabolism as a key pathway for TL regulation

Utilising the prioritised gene list as well as closest genes to the sentinel variants, we showed several pathways were enriched for telomere-associated loci. Of note, we observed significant overrepresentation of genes in several nucleotide metabolism pathways (Supplementary Table 11 and Figure 2.2). Key genes highlighted within these pathways involve in both the biosynthesis (*TYMS*, *TK1* and *DCK*) and catabolism (*SAMHD1*) of dNTPs. Biosynthesis of dNTPs occurs via two routes, *de-novo* synthesis and nucleotide salvage pathway. TK1 and DCK are the rate-limiting enzymes that catalyse the first step of the salvage pathway of nucleotide biosynthesis, converting deoxynucleosides to their monophosphate forms (dNMPs) before other enzymes facilitate further phosphorylation to dNDPs and dNTPs (Figure 2.2B)[273–276]. TYMS is considered as a component of the *de novo* pathway, and is the key regulator of dTMP biosynthesis, converting dUMP to dTMP[277,278]. However, as dUMP substrates can be derived from either *de novo* synthesis or deamination of dCMP produced from the salvage pathway, it could be considered to function within both pathways (Figure 2.2B)[276]. Besides controlling biosynthetic pathways, equilibrium of cellular dNTP levels is also

achieved by regulating degradation of dNTPs, a key regulator of which is the SAMHD1. It catalyses hydrolysis of dNTPs to deoxynucleosides and triphosphates, thereby preventing accumulation of excess dNTPs. Besides fine-tuning the dNTP supply system, potential errors in base insertion into DNA synthesis are monitored by another prioritised enzyme, the base excision repair enzyme, SMUG1[248,251].

A balanced cellular pool of dNTPs is required for DNA replication and repair as well as maintaining proliferative capacity and genome stability. Low levels of dNTPs can induce replication stress, subsequently leading to increased mutation rates[249,279]. A surplus of dNTPs, on the other hand, reduces replication fidelity, which also causes higher levels of spontaneous mutagenesis[280,281]. A dynamic balance between biosynthesis and catabolism is required to maintain the equilibrium. Since maintaining the balance of intracellular dNTP pool is also fundamental to other pathways that are implicated in telomere homeostasis, including cellular proliferation and DNA repair, disruption of dNTP homeostasis may trigger a sequence of cellular events that interplay synergistically leading to abnormalities of TL and genome instability.

By clustering the prioritised genes via their functional connections, we highlighted a number of pathways that were enriched for TL regulation, which included DNA replication, transcription and repair, cell cycle regulation, immune response and intracellular trafficking (Figure 2.2A). However, we noted that because the gene prioritisation was based on integration of bioinformatic evidence from several publicly available databases, which also laid the foundation for establishing pathways used in the enrichment analyses, this approach may suffer from self-fulfilling circular arguments. For example, genes involved in telomere homeostasis have been prioritised as likely-causal genes, such as *TERT*, *TERC*, *TERF2*, *RTEL1* and *POT1*, because they have well-established roles in telomere maintenance; however, in pathway enrichment analysis, they are also the key contributors that drive the telomere-related pathways that were significantly enriched, such as regulation of telomere maintenance via telomere lengthening, regulation of telomeric loop disassembly, telomere capping etc. (Supplementary Table 11). Nevertheless, through analysing functional interconnections between our prioritised gene candidates, we can highlight some of the known as well as novel mechanisms underlying TL regulation, and provide more relevant gene targets for future experimental follow-up studies.

Chapter 3

Clinical relevance of LTL to cardiometabolic and other common, chronic conditions

Abstract

Background LTL, as a marker of biological age, has been associated with various age-related diseases, including CVD and cancers. However, the degree to which LTL is causally linked to these diseases is uncertain, and no systematic studies for causal roles of LTL in a broad spectrum of common chronic diseases have been conducted.

Objectives To refine previously reported and discover novel associations of LTL with cardiometabolic traits and diseases, as well as other common chronic conditions, including cancers.

Methods Observational association between LTL and incident T2D risk was analysed in EPIC-InterAct case-cohort study using Prentice-weighted Cox proportional hazards regression models with different covariate adjustments. Causal associations with cardiometabolic diseases and risk factors were analysed using summary-level results under the MR framework with genetic instruments at genome-wide significance or FDR<0.05. Sensitivity analyses were performed by excluding the *HLA* region or implementing other MR models. Phenome-wide association study was performed in >350,000 UK Biobank participants testing causal associations of LTL with a broad range of common diseases using allele score MR method.

Results LTL was not associated with T2D risk, either observationally or genetically determined. Genetically-longer leukocyte telomeres were associated with lower risk for CHD, as previously suggested, but only when using genome-wide significant (OR[95%CI] = 0.87[0.80-0.94], *p*-value=$4.42X10^{-4}$), not FDR variants (OR[95%CI] = 1.05[0.99-1.10], *p*-value=0.08). They were also associated with higher levels of established cardiovascular risk factors, including blood pressure, adiposity and circulatory lipid levels; and a range of proliferative conditions, including both malignant as well as non-malignant neoplasms, in the phenome-wide analyses.

Conclusions Our analyses refine previously reported clinical relevance of LTL, including relevance to cardiometabolic traits and diseases as well as cancers, and systematically characterise potential roles of telomere dysregulation in a broad spectrum of human diseases, deepening our understanding of aetiological implication of LTL in these diseases.

3.1 Introduction

Severe telomere loss, through loss of function mutations of core telomere and telomerase components leads to several diseases, which share features such as bone marrow failure and organ damage. These "telomere syndromes" include rare, childhood-onset diseases such as dyskeratosis congenita, and adult-onset diseases including aplastic anaemia and idiopathic pulmonary fibrosis[172–174,282–284]. One of the common features of the telomere syndromes is premature ageing, as outlined in the introduction (section 1.4.1.4.1). Together with shorter TL observed at older ages, this has led to TL (most commonly measured in human leukocytes (LTL)) to be proposed as a marker of biological age. Observational studies have linked LTL to risks of a range of common age-related diseases, including CAD and some cancers[285–290]. However, the degree to which LTL is causally linked to these diseases is unknown as the observed associations may have been due to reverse causation or confounding. Previous MR analyses using independent variants in up to 10 genome-wide significant loci as genetic instruments for LTL have found some causal evidence to support such associations[158,177,179,291], but power of the earlier MR studies has been limited by the number and strength of loci identified and sample sizes of outcome GWAS. Comparison to observational estimates for selected disease outcomes, such as various cancers and CVD, has suggested discrepant findings where certain cancers exhibited strong evidence of associations in MR but completely no evidence in observational analyses, such as lung cancer, melanoma and glioma cancers[158], or the patterns and directions of associations were different in MR and observational analyses (section 1.4.1.4.2). Whether the MR results were driven by specific loci or biased by horizontal pleiotropy or the observational results influenced by confounding and reverse causality are unclear. In addition, associations of genetically determined LTL with disease outcomes have not been performed systematically to elucidate potential causal implications of LTL in a broad spectrum of common chronic diseases.

In this chapter, I will characterise roles of LTL in various types of human diseases, including cardiometabolic and other common chronic diseases. Using the expanded pool of genetic instruments that increased total LTL variability explained, with over 30% or 50% of total chip-based heritability of LTL (h^2=5.0%) explained by independent variants at genome-wide significance or FDR<0.05, respectively, I substantially increased power of MR studies.

Moreover, with largely increased spectrum of clinical outcomes from biobank cohort studies, I discovered causal relevance of LTL to more clinical phenotypes that have not been previously studied, such as benign tumours and non-neoplastic, common chronic diseases. With increased sample sizes from consortia-led GWAS meta-analyses, I refined uncertain associations, such as those with cardiometabolic diseases and related risk factors. Overall this chapter identifies more clinical outcomes linked to and refines uncertain associations with genetically determined LTL, shedding light upon aetiological mechanisms that underlie telomere dysregulation-related diseases.

3.2 Methods

3.2.1 Observational association of LTL with T2D

Observational association between LTL and incident T2D was analysed in EPIC-InterAct, a large-scale prospective case-cohort study, described in more detail previously and in other sections (sections 2.2.1 and 5.2.1.1, Supplementary Notes)[212,213]. The association was analysed separately in each country, using Prentice-weighted Cox proportional hazards regression models with age as the underlying timescale[212,292] and different adjustments for confounding. The basic model included adjustments for age, sex, centre and batch; other covariates were added successively, including Mediterranean diet score, lifetime pattern for alcohol consumption frequency, smoking intensity, questionnaire-based physical activity index, the highest level of education attained, BMI and waist circumference. Moreover, a multivariable model that included all covariates except for two measures of adiposity, BMI and waist circumference; and two models each with BMI or waist circumference were tested. These multivariable models were designed to take account of more potential confounders and examine effects of obesity on the association of LTL with T2D. Continuous variables, except age, were normalised via inverse normal transformation in each country separately; categorical variables were coded ordinally. The resultant hazard ratios (HRs) were meta-analysed across countries using random-effects meta-analysis models.

3.2.2 Assessment of causal effects of LTL on cardio-metabolic traits and diseases

3.2.2.1 Cardio-metabolic diseases

Causal effects of LTL on the risks of T2D and CHD were studied using summary-level results under the MR framework. Two sets of genetic instruments were defined, one with conditionally independent lead SNPs of loci at genome-wide significance (p-value=$5X10^{-8}$), and the other more inclusive with all variants that reached the FDR threshold of 0.05. Association estimates with the exposure (LTL) were taken from the final GWAS meta-analysis (chapter 2), and applied to recent large-scale GWAS meta-analyses for T2D and CVD: for T2D, this consisted of three non-overlapping studies/consortia (EPIC-InterAct, UK Biobank, and

DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) v3[293] with exclusion of the InterAct samples), for a total number of 44,417 cases and 489,910 controls; for CHD, this consisted of a meta-analysis of results from CARDIoGRAMplusC4D (Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics) and UK Biobank, for a sample size of 85,358 cases and 551,249 controls.

In the primary analysis, I first calculated the Wald ratio (causal estimate) for each individual genetic instrument using the two-sample MR method, ($\frac{\widehat{\beta_{Y_k}}}{\widehat{\beta_{X_k}}}$, SE=$\frac{\sigma_{Y_k}}{\widehat{\beta_{X_k}}}$; β_Y, β_X, beta coefficients for associations with disease outcomes (T2D or CHD) and LTL, respectively; *k* = 1,2…the total number of genetic instruments), and then pooled the ratio estimates using inverse-variance weighted fixed-effects meta-analysis models, mathematically equivalent to weighted linear regression models forced through the origin. In the secondary sensitivity analyses, I applied alternative MR methods to correct for horizontal pleiotropic outliers, including the MR Egger regression method[68,294] and the (penalised) weighted median MR method[74]. I also removed the HLA region due to its high pleiotropy, from both sets of the genetic instruments and re-applied the same MR methods with HLA-excluded genetic instruments.

3.2.2.2 Cardio-metabolic traits

I examined associations between genetically predicted longer LTL and a series of continuous traits associated with CVD risk, including glycaemic, lipid, blood pressure and adiposity-related measures. Summary statistics for associations of genetic instruments with glycaemic traits were retrieved from the MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium)[295,296], with lipid traits from the GLGC (Global Lipids Genetics Consortium)[297], with blood pressure traits from the UK biobank and with adiposity-related traits from a meta-analysis of the GIANT (Genetic Investigation of ANthropometric Traits) and UK Biobank studies[298,299].

I then utilised the same MR approaches as described above in the 3.2.2.1, including approaches used in both the primary and secondary analyses. For genetic instruments that were missing in the outcome meta-analysis results, I selected proxy SNPs that were genetically correlated (LD r^2>0.8) and physically closest to them if available.

3.2.3 Genetic correlations of LTL to human phenotypes and ageing-related traits

Cross-trait LD score regression (LDSC) analysis was used to measure genetic correlations between LTL and selected traits within the LD Hub database[300,301] (version 1.4.1). From 832 available traits curated in the LD Hub, I filtered them for the purpose of QC and avoiding redundancy. To be more specific, I removed traits and diseases without prior evidence of genetic bases (heritability z-score < 2), traits of medication uses, lipid sub-fractions and for which studies with sample sizes<1000. Where multiple datasets of a same trait existed, I firstly prioritised results from the largest or most recent consortia-led studies over UK Biobank samples only studies. Following this, I prioritised the trait selection based on sample size, (publication) date and quality of outcome definition (diagnosed conditions versus self-reported only). In total, there were 320 phenotypes included in the analysis, covering diverse conditions ranging from behavioural risk factors to common complex diseases.

Genome-wide summary statistics were used as inputs. In addition to standardised QC implemented within the software, variants with MAF (<1% for HapMap3 or <5% for 1000 Genomes EUR imputed SNPs), sample size (<0.67 times 90th percentile of variants' sample sizes), alleles not aligned to 1000 Genomes, or insertions or deletions or structural variants were removed.

3.2.4 Phenome-wide association study (PheWAS)

3.2.4.1 UK Biobank

The UK Biobank study is a population-based cohort of 500,000 people aged between 40 and 69 years and recruited from multiple centres in UK from 2006 to 2010[302]. A range of modifiable factors were taken via questionnaires and nurse interviews (such as demographic features, family histories of diseases, health status and lifestyles); anthropometric measurements, blood pressures and circulatory biomarkers were measured; and blood, urine and saliva samples were taken for future analyses. Genome-wide genetic data has been collected for every participant with purpose-designed genotyping arrays (the UK Biobank BiLEVE and Axiom Arrays) and processed with extensive QC procedures. Clinical follow-up

data has been provided through linkage to health and medical records, including primary care data and data from national hospital data electronic record systems and national death and cancer registries. Detailed description of these datasets can be found elsewhere[303]. In this study, participants that are in close familial relationships (equal to or closer than the third degree), or of non-European descent were excluded, resulting in 352,071 individuals for the PheWAS.

3.2.4.1 PheWAS on manually curated clinical outcomes

Using two-sample MR methods[304,305] we investigated potential causal effects of LTL on 122 diseases manually curated in UK Biobank[306] (Supplementary Table 13). Disease definitions were generated using self-reported histories of diseases or disease-relevant medical treatments, combined with records of hospitalisation for the 9th/10th revision of the WHO International Classification of Diseases (ICD9/10)-coded clinical outcomes. Diseases were selected where there were sufficient case numbers to have 80% power to detect an odds ratio (OR) of 1.1 or 0.9 at the 5% alpha level (Supplementary Table 14). LTL was genetically proxied by 52 independently associated variants (FDR<0.05). In addition, individual SNP effects on diseases were tested using logistic regression in SNPTEST[307] adjusted for sex, age, genotyping array and the top 5 genetic PCs. Causal association estimates were calculated using the inverse variance weighted MR approach. Sensitivity analyses were performed using median-based MR[308], MR-RAPS (Mendelian Randomization using Robust Adjusted Profile Score)[309] and MR-Egger regression[310] to cross-validate results and evaluate unbalanced pleiotropic effects.

3.2.4.2 PheWAS on the full set of ICD10-codes defined clinical outcomes

To analyse associations between genetically predicted longer LTL and a broad spectrum of clinical outcomes, we performed PheWAS on all individual ICD10-coded diseases at level 2. The analyses were restricted to diseases with case numbers greater than 500. In addition, I also analysed 35 self-reported cancers, among which 27 were combined with corresponding ICD10-coded cancer diagnoses (Supplementary Table 16). Logistic regression was used with ICD10-defined outcome cases coded as 1, and the rest as controls coded as 0. Adjustments included age, sex, genotyping array and the top 10 PCs. For each participant, we constructed

PRS of LTL (alleles aligned to the direction of increasing LTL) by summing up weighted dosages of conditionally independent genome-wide significant SNPs. SNP weights were defined as absolute values of association beta coefficient estimates.

I used the glm() function implemented in R with binomial distribution to test associations of weighted PRS of LTL with each of the disease outcomes. Moreover, I adapted the PheWAS R package[311], converting ICD9 to ICD10 codes, and examined associations between disease outcomes and the weighted PRS of LTL, as well as individual locus sentinel SNPs at genome-wide significance. The same database of disease outcomes was used for the glm() function and the PheWAS R package based methods. The two methods produced exactly the same results for overlapping associations tested. Associations with *p*-values smaller than $1.3x10^{-4}$ (Bonferroni corrected for a total number of 370 diseases tested) were reported as being statistically significant.

3.2.5 Variants-based cross-database query

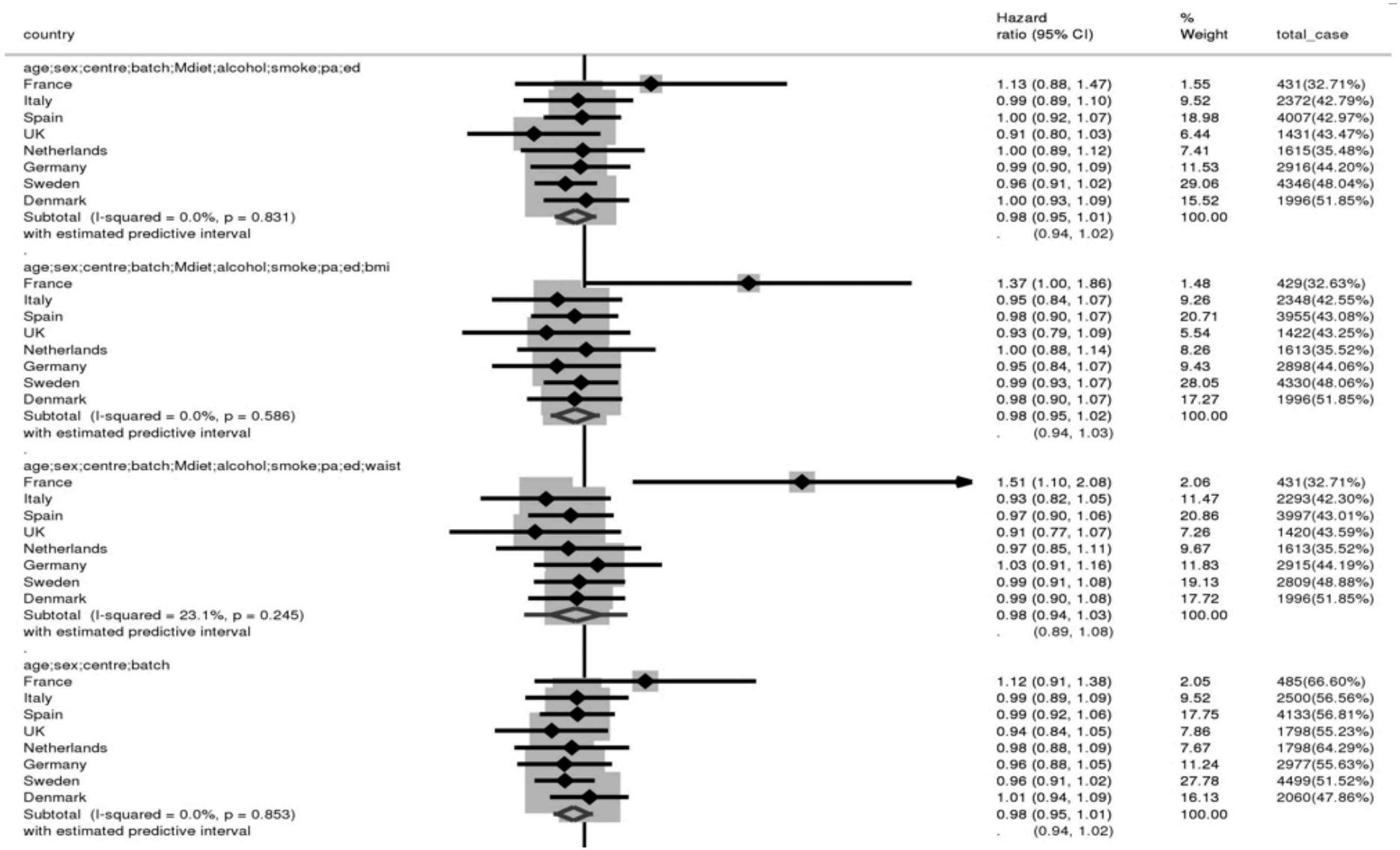
Independent variants (FDR<0.05) and their strong proxies ($r^2 \geq 0.8$) were queried against publicly available GWAS results using PhenoScanner[312] for computational efficiency. A list of GWAS results implemented in the software was previously published[312]. Results were filtered to include associations with *p*-values<$1X10^{-6}$ and in high LD ($r^2 \geq 0.6$) with locus sentinel SNPs. To avoid redundancy, for each trait, only results from the most recent and/or largest studies were retained.

3.3 Results

3.3.1 Observational association between LTL and incident T2D

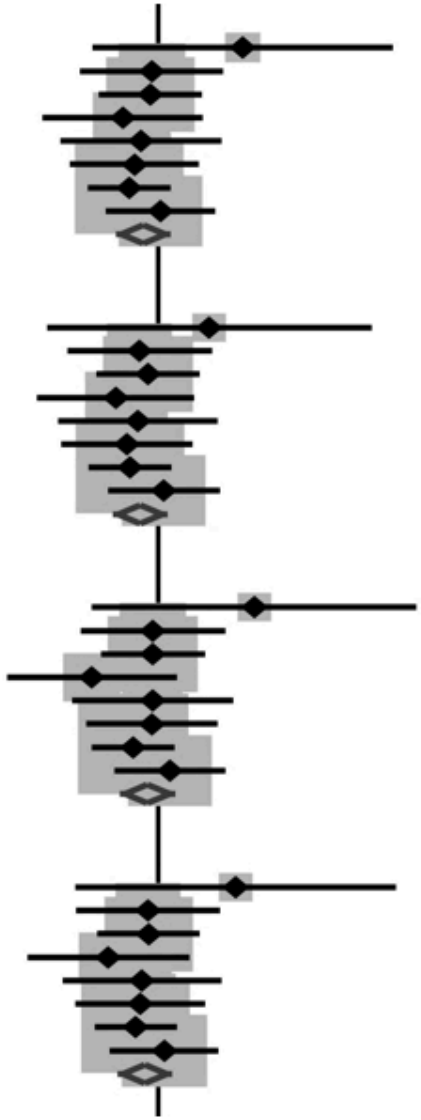
Using the largest incident T2D case-cohort study (EPIC-InterAct study, n = 29,238; 45.1% cases), I investigated whether measured differences in LTL were associated with development of future T2D. No evidence has been found to support an association between LTL and incident T2D in any of the models (Figure 3.1) and no substantial between-country heterogeneity was found (I^2<30% in any models). Of note, France showed a distinct pattern of associations compared to the other countries across all models tested, especially after adjusting for adiposity-related traits (BMI or waist circumference), such that French participants (n=485, all women) showed larger, but opposite associations. However, due to the small sample size of the French subset, confidence intervals were large (HR[95%CI] = 1.12[0.91-1.38] from the basic model and 1.51[1.10-2.08] from a multivariable model adjusted for waist circumference) and the overall contribution of the French subset to the meta-analysis was small (weight=1.48-2.13% across models).

Figure 3.1. Observational association between LTL and incident T2D risk.
X-axis indicates HRs of T2D per 1-SD increase of LTL. Alcohol_re: alcohol consumption frequency; Mdiet: Mediterranean diet score; pa: physical activity; ed: the highest educational level; bmi: body mass index; waist: waist circumference



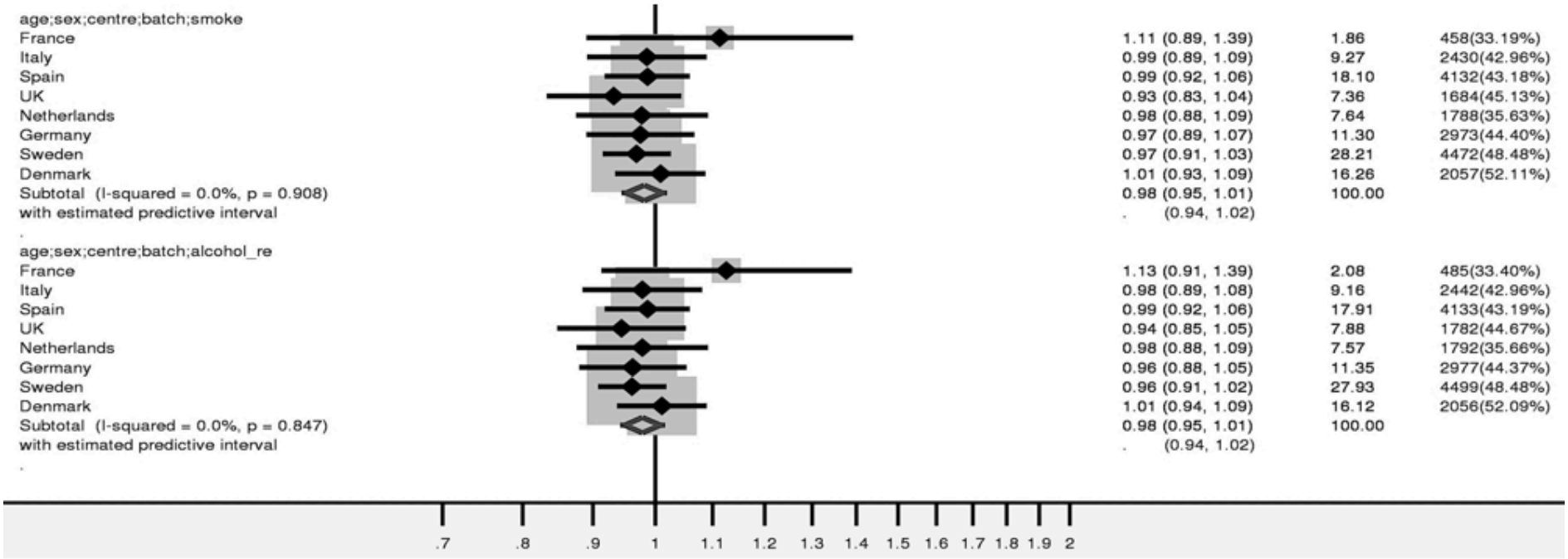
country	Hazard ratio (95% CI)	% Weight	total_case
age;sex;centre;batch;Mdiet;alcohol;smoke;pa;ed			
France	1.13 (0.88, 1.47)	1.55	431(32.71%)
Italy	0.99 (0.89, 1.10)	9.52	2372(42.79%)
Spain	1.00 (0.92, 1.07)	18.98	4007(42.97%)
UK	0.91 (0.80, 1.03)	6.44	1431(43.47%)
Netherlands	1.00 (0.89, 1.12)	7.41	1615(35.48%)
Germany	0.99 (0.90, 1.09)	11.53	2916(44.20%)
Sweden	0.96 (0.91, 1.02)	29.06	4346(48.04%)
Denmark	1.00 (0.93, 1.09)	15.52	1996(51.85%)
Subtotal (I-squared = 0.0%, p = 0.831)	0.98 (0.95, 1.01)	100.00	
with estimated predictive interval	. (0.94, 1.02)		
.			
age;sex;centre;batch;Mdiet;alcohol;smoke;pa;ed;bmi			
France	1.37 (1.00, 1.86)	1.48	429(32.63%)
Italy	0.95 (0.84, 1.07)	9.26	2348(42.55%)
Spain	0.98 (0.90, 1.07)	20.71	3955(43.08%)
UK	0.93 (0.79, 1.09)	5.54	1422(43.25%)
Netherlands	1.00 (0.88, 1.14)	8.26	1613(35.52%)
Germany	0.95 (0.84, 1.07)	9.43	2898(44.06%)
Sweden	0.99 (0.93, 1.07)	28.05	4330(48.06%)
Denmark	0.98 (0.90, 1.07)	17.27	1996(51.85%)
Subtotal (I-squared = 0.0%, p = 0.586)	0.98 (0.95, 1.02)	100.00	
with estimated predictive interval	. (0.94, 1.03)		
.			
age;sex;centre;batch;Mdiet;alcohol;smoke;pa;ed;waist			
France	1.51 (1.10, 2.08)	2.06	431(32.71%)
Italy	0.93 (0.82, 1.05)	11.47	2293(42.30%)
Spain	0.97 (0.90, 1.06)	20.86	3997(43.01%)
UK	0.91 (0.77, 1.07)	7.26	1420(43.59%)
Netherlands	0.97 (0.85, 1.11)	9.67	1613(35.52%)
Germany	1.03 (0.91, 1.16)	11.83	2915(44.19%)
Sweden	0.99 (0.91, 1.08)	19.13	2809(48.88%)
Denmark	0.99 (0.90, 1.08)	17.72	1996(51.85%)
Subtotal (I-squared = 23.1%, p = 0.245)	0.98 (0.94, 1.03)	100.00	
with estimated predictive interval	. (0.89, 1.08)		
.			
age;sex;centre;batch			
France	1.12 (0.91, 1.38)	2.05	485(66.60%)
Italy	0.99 (0.89, 1.09)	9.52	2500(56.56%)
Spain	0.99 (0.92, 1.06)	17.75	4133(56.81%)
UK	0.94 (0.84, 1.05)	7.86	1798(55.23%)
Netherlands	0.98 (0.88, 1.09)	7.67	1798(64.29%)
Germany	0.96 (0.88, 1.05)	11.24	2977(55.63%)
Sweden	0.96 (0.91, 1.02)	27.78	4499(51.52%)
Denmark	1.01 (0.94, 1.09)	16.13	2060(47.86%)
Subtotal (I-squared = 0.0%, p = 0.853)	0.98 (0.95, 1.01)	100.00	
with estimated predictive interval	. (0.94, 1.02)		

to be continued on the next page



.			
age;sex;centre;batch;pa			
France	1.12 (0.91, 1.39)	2.13	485(33.40%)
Italy	0.99 (0.90, 1.09)	9.34	2439(42.89%)
Spain	0.99 (0.92, 1.06)	18.12	4133(43.19%)
UK	0.95 (0.85, 1.06)	7.50	1691(45.36%)
Netherlands	0.98 (0.87, 1.09)	7.36	1667(35.45%)
Germany	0.97 (0.88, 1.06)	11.49	2977(44.37%)
Sweden	0.96 (0.91, 1.02)	27.98	4483(48.41%)
Denmark	1.00 (0.93, 1.08)	16.08	2059(52.16%)
Subtotal (I-squared = 0.0%, p = 0.893)	0.98 (0.95, 1.01)	100.00	
with estimated predictive interval	. (0.94, 1.02)		
.			
age;sex;centre;batch;Mdiet			
France	1.07 (0.86, 1.35)	1.84	475(33.05%)
Italy	0.97 (0.88, 1.08)	9.26	2388(42.76%)
Spain	0.99 (0.92, 1.06)	18.25	4037(43.05%)
UK	0.94 (0.84, 1.05)	7.87	1740(44.25%)
Netherlands	0.97 (0.87, 1.09)	7.63	1750(35.66%)
Germany	0.96 (0.87, 1.05)	11.21	2920(44.18%)
Sweden	0.96 (0.91, 1.02)	28.24	4396(48.16%)
Denmark	1.01 (0.93, 1.09)	15.71	2002(51.85%)
Subtotal (I-squared = 0.0%, p = 0.945)	0.97 (0.95, 1.01)	100.00	
with estimated predictive interval	. (0.94, 1.01)		
.			
age;sex;centre;batch;ed			
France	1.14 (0.91, 1.43)	1.87	466(33.26%)
Italy	0.99 (0.90, 1.10)	9.42	2437(42.88%)
Spain	0.99 (0.92, 1.07)	18.25	4104(43.13%)
UK	0.91 (0.81, 1.03)	6.86	1517(44.17%)
Netherlands	0.99 (0.89, 1.11)	7.65	1781(35.54%)
Germany	0.99 (0.90, 1.09)	11.43	2977(44.37%)
Sweden	0.97 (0.91, 1.02)	28.64	4485(48.41%)
Denmark	1.02 (0.94, 1.10)	15.87	2058(52.14%)
Subtotal (I-squared = 0.0%, p = 0.724)	0.98 (0.95, 1.02)	100.00	
with estimated predictive interval	. (0.95, 1.02)		
.			
age;sex;centre;batch;smoke			
France	1.11 (0.89, 1.39)	1.86	458(33.19%)
Italy	0.99 (0.89, 1.09)	9.27	2430(42.96%)
Spain	0.99 (0.92, 1.06)	18.10	4132(43.18%)
UK	0.93 (0.83, 1.04)	7.36	1684(45.13%)
Netherlands	0.98 (0.88, 1.09)	7.64	1788(35.63%)
Germany	0.97 (0.89, 1.07)	11.30	2973(44.40%)
Sweden	0.97 (0.91, 1.03)	28.21	4472(48.48%)
Denmark	1.01 (0.93, 1.09)	16.26	2057(52.11%)
Subtotal (I-squared = 0.0%, p = 0.908)	0.98 (0.95, 1.01)	100.00	
with estimated predictive interval	. (0.94, 1.02)		

to be continued on the next page


age;sex;centre;batch;smoke
France
Italy
Spain
UK
Netherlands
Germany
Sweden
Denmark
Subtotal (I-squared = 0.0%, p = 0.908)
with estimated predictive interval
.
age;sex;centre;batch;alcohol_re
France
Italy
Spain
UK
Netherlands
Germany
Sweden
Denmark
Subtotal (I-squared = 0.0%, p = 0.847)
with estimated predictive interval
.
1.11 (0.89, 1.39) 1.86 458(33.19%)
0.99 (0.89, 1.09) 9.27 2430(42.96%)
0.99 (0.92, 1.06) 18.10 4132(43.18%)
0.93 (0.83, 1.04) 7.36 1684(45.13%)
0.98 (0.88, 1.09) 7.64 1788(35.63%)
0.97 (0.89, 1.07) 11.30 2973(44.40%)
0.97 (0.91, 1.03) 28.21 4472(48.48%)
1.01 (0.93, 1.09) 16.26 2057(52.11%)
0.98 (0.95, 1.01) 100.00
. (0.94, 1.02)
1.13 (0.91, 1.39) 2.08 485(33.40%)
0.98 (0.89, 1.08) 9.16 2442(42.96%)
0.99 (0.92, 1.06) 17.91 4133(43.19%)
0.94 (0.85, 1.05) 7.88 1782(44.67%)
0.98 (0.88, 1.09) 7.57 1792(35.66%)
0.96 (0.88, 1.05) 11.35 2977(44.37%)
0.96 (0.91, 1.02) 27.93 4499(48.48%)
1.01 (0.94, 1.09) 16.12 2056(52.09%)
0.98 (0.95, 1.01) 100.00
. (0.94, 1.02)
.7 .8 .9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2

3.3.2 Associations of genetic differences in LTL with cardio-metabolic diseases and traits

3.3.2.1 T2D

Although observational analyses provided no evidence of a significant or strong association between LTL and incident T2D, a negative confounding effect cannot be ruled out. When such confounding effects are not properly adjusted, the observed association can be biased towards null. To minimise effects from unadjusted confounders and reverse causation, summary-level, two-sample MR approaches were used. I observed no significant associations between genetically determined LTL and T2D (OR[95%CI] = 1.00[0.92-1.10] for 1-SD increase of genetically predicted LTL, *p*-value=0.92, Table 3.1 and 3.2A) using conditionally independent genome-wide significant variants as genetic instruments. The intercept of the MR-Egger regression model was not significantly different from 0 (*p*-value=0.15, Table 3.2A), indicating there was no statistically significant directional pleiotropy detected, although there was a moderate heterogeneity observed (I^2=55.7%). The effect size in the MR Egger regression model was smaller (OR[95%CI] = 0.82[0.66-1.01], *p*-value=0.07, Table 3.2A), yet below nominal significance.

Using the inverse-variance weighted MR approach but FDR variants as genetic instruments, a nominally significant association was observed (OR[95%CI] = 0.94[0.93-0.99], *p*-value=0.01, Table 3.2B), yet below the Bonferroni corrected threshold (*p*-value=2.08X10^{-3}, assuming 24 traits/diseases as independent tests). The between-variant heterogeneity was slightly larger among the FDR variants (I^2=62.6%).

HLA was shown to be strongly associated with the risk of T2D (OR[95%CI] = 2.51[1.60-3.93]), and because this region has been known to be highly pleiotropic, which might violate the MR assumption that genetic instruments cannot be associated with any confounders, and not with the outcome conditioning on the exposure, we performed sensitivity analyses excluding the *HLA* region. These showed that the *HLA* did not influence the MR results, as effect sizes were not significant in either stringent or more inclusive (genome-wide versus FDR variants) models (OR[95%CI] = 0.97[0.88-1.06], *p*-value=0.51 and OR[95%CI] = 0.95[0.90-1.01], *p*-value=0.08, respectively).

3.3.2.2 CHD

Shorter LTL has been widely acknowledged to be associated with a higher risk of CHD[158,177]. Using our new genetic instruments that explained a larger proportion of the variability of LTL, I replicated earlier studies that showed genetic predisposition to longer LTL was associated with a lower risk of CHD (Table 3.1 and 3.2). One SD increase of genetically predicted LTL was associated with about 23% reduced CHD risk using the inverse variance weighted MR method with conditionally independent genome-wide significant variants as instrumental variables (OR[95%CI] = 0.87[0.80-0.94], *p*-value=4.42X10^{-4}, Table 3.1 and 3.2A). The protective effect of longer LTL was in line with previous findings where 7 previously identified genome-wide significant variants[151] were used as instrumental variables to assess causalities of LTL on CAD[151] and CVD[177], in CARDIoGRAM and UK biobank interim release datasets, respectively. The effect sizes were comparable, yet the association strength in our analysis was larger possibly due to the better genetic instruments used and the larger sample size of the outcome GWAS dataset.

The result was further validated in sensitivity analyses using the MR Egger regression method (OR[95%CI] = 0.70[0.58-0.85], *p*-value=2.42X10^{-4}, Table 3.2A). The Egger intercept term was not significantly different from 0 (*p*-value=0.24, Table 3.2A), suggesting no statistically detectable horizontal pleiotropy among the genome-wide significant variants, even though the overall heterogeneity of causal estimates among these genome-wide significant variants was substantial (I^2=76.9%). In addition, *HLA* region showed a strong causal effect on CHD (Figure 3.2), but in the opposite direction compared to the overall effect. Exclusion of *HLA* improved the significance of the MR association (OR[95%CI] = 0.83[0.76-0.90], *p*-value=1.56X10^{-5}), and slightly reduced the overall heterogeneity (I^2=60.9%).

Moreover, using the extended list of variants at FDR loci, the protective effect of genetically determined longer LTL on CHD became much weaker and below nominal significance (OR[95%CI] = 1.05[0.99-1.10], *p*-value=0.08, Table 3.2B). Exclusion of the *HLA* region did not improve the association strength, as a small weight (1.76%) was given to the *HLA* region. The association was even weaker and remained to be non-significant using the MR-Egger approach (OR[95%CI] = 1.01[0.91-1.15], *p*-value=0.90). The Egger intercept term was statistically no different from 0 (*p*-value=0.52), indicating there was no evidence for unbalanced pleiotropy among the FDR variants. Given that FDR variants almost doubled the variation explained by genome-wide significant variants, and with these two different sets of

genetic instruments, the same dataset for CHD summary statistics was used, such discrepant results might suggest that there were specific loci that drove the causal association observed between LTL and CHD. This notion was supported by further analyses on causalities of individual variants.

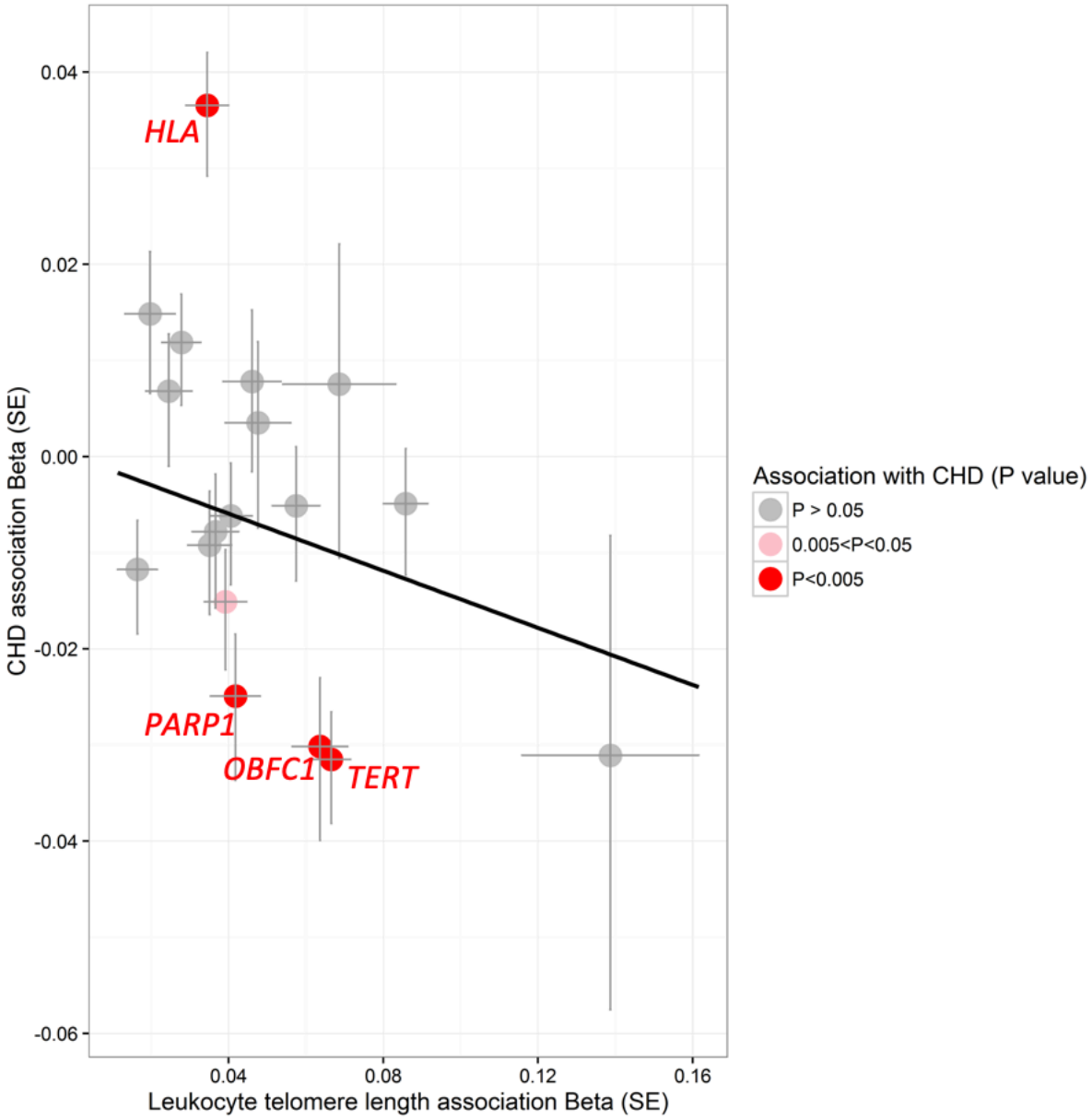
Further analyses of individual genome-wide significant variants showed that lead variants at specific loci (*PARP1*, *TERT* and *OBFC1* loci) were strong drivers of the protective effect of genetically longer LTL on CHD, due to their relatively larger and more precise estimates of the causal effects (Figure 3.2), with all being directionally consistent with the overall effect.

Table 3.1 Associations between genetically predicted LTL and cardio-metabolic diseases.
The associations with T2D and CHD are presented in each study that contributed to the outcome GWAS meta-analysis separately and overall. LTL was predicted using sentinel variants at genome-wide significant loci.

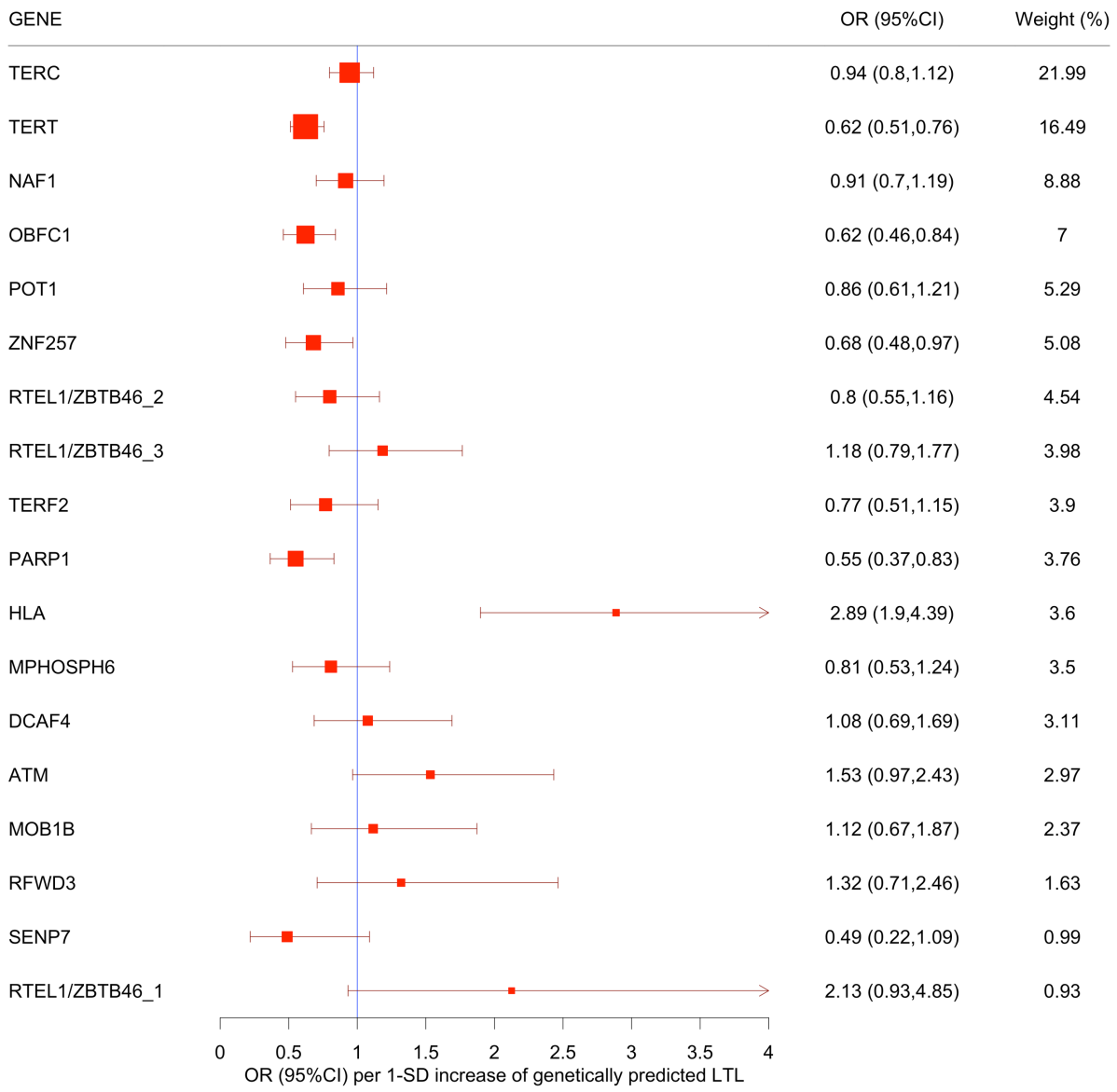
Disease	Study	OR [95% CI]	weight (%)	case N	control N	*p*-value
	UK Biobank	0.83 [0.74-0.93]	52.17	24,557	427,745	$9.29X10^{-4}$
CHD	CARDioGRAMplusC4D	0.91 [0.81-1.02]	47.83	60,801	123,504	0.10
	Overall	0.87 [0.80-0.94]	100	85,358	551,249	$4.42X10^{-4}$
	UK Biobank	1.08 [0.97-1.21]	68.28	25,529	424,577	0.15
	EPIC-InterAct CoreExome chip	0.86 [0.62-1.21]	7.28	5,121	7,269	0.39
T2D	EPIC-InterAct 660W chip	0.83 [0.56-1.23]	5.2	4,187	4,254	0.36
	DIAGRAM	0.86 [0.70-1.05]	19.24	34,840	114,981	0.14
	Overall	1.00 [0.92-1.10]	100	69,677	551,081	0.92

Figure 3.2 Pleiotropic effects of LTL-associated variants (p-value<5X10^{-8}) on CHD risk.
Association beta coefficients are plotted with bars indicating standard errors (SE). Colours indicate association strength (p-values) for each variant with the CHD risk, as shown in the legend. None of the variants reached genome-wide significance level for their associations with the CHD risk.

A. Scatter plot showing variants' effect sizes on LTL *versus* CHD.



B. Forest plot showing causal effect estimate for each individual locus sentinel variant. Variants were labelled with candidate genes for the corresponding loci, and ordered by their weights.



3.3.2.3 Cardio-metabolic traits

We systematically analysed associations of genetically predicted increase of LTL with various cardio-metabolic traits. These traits were categorised into several groups, including lipid, glycaemic, blood pressure and adiposity-related traits. Genetically predicted longer LTL was found to be significantly associated with higher levels of diastolic and systolic blood pressures, body fat percentage, BMI-adjusted waist hip ratio, total triglyceride and cholesterol and LDL cholesterol levels (Table 3.2A). The BMI-adjusted waist and hip circumferences were associated with genetically predicted LTL only in the MR-Egger regression, but not the inverse variance weighted MR analyses, possibly due to biases from unbalanced pleiotropy in the latter (Egger intercept term *p*-values = 0.09 and 0.06, respectively, Table 3.2A).

Results were similar when using the FDR variants (Table 3.2B). However, because these FDR variants-based MR analyses were restricted to traits for which summary results of their recent GWAS were available and conducted within Europeans and imputation panels were up to date, fewer traits were analysed.

Table 3.2 Associations between genetically predicted LTL and cardio-metabolic diseases and traits.
LTL was predicted using **A**. conditionally independent genome-wide significant variants or **B**. FDR variants.

A

Independent variable	Dependent variable		Inverse variance weighted MR method		Heterogeneity test	
			Beta (SE) for continous variable	*p*-value	Cochran's Q test	
	Category	Phenotype	OR [95% CI] for binary variable		Cochran's Q	*p*-value
One-SD increase of genetically predicted LTL	Disease	CHD	0.87 [0.80-0.94]	4.42E-04	15.48	0.56
		T2D	1.00 [0.92-1.10]	0.92	9.40	0.93
	Adiposity	BMI	-0.02 (0.01)	0.03	4.33	1.00
		BMI-adjusted Wasit/Hip Ratio	0.06 (0.01)	3.51E-08	7.06	0.98
		BMI-adjusted Wasit Circumference	-0.01 (0.01)	0.56	7.79	0.97
		BMI-adjusted Hip Circumference	-0.02 (0.01)	0.17	9.71	0.92
		Body Fat Percentage	-0.04 (0.01)	2.85E-03	11.32	0.84
	Lipid	Total triacylglycerol	0.08 (0.03)	2.77E-03	3.58	1.00
		Total cholesterol	0.11 (0.03)	3.93E-04	5.13	1.00
		LDL	0.09 (0.03)	2.85E-03	5.65	1.00
		HDL	-0.03 (0.03)	0.34	5.48	1.00
	Blood Pressure	Diastolic BP	0.06 (0.01)	7.33E-07	9.69	0.92
		Systolic BP	0.12 (0.01)	3.70E-22	6.67	0.99
	Glycaemia	Insulin Sensitivity Index Adjusted for BMI	0.01 (0.06)	0.90	6.07	0.99
		Insulin Secretion at 30min during OGTT	-0.02 (0.14)	0.89	12.20	0.79
		HbA_{1C}	0.02 (0.02)	0.26	1.14	1.00
		Fasting Plasma Glucose	0.02 (0.01)	0.22	1.23	1.00
		BMI-adjusted Fasting Plasma Glucose	0.01 (0.01)	0.31	1.21	1.00
		Fasting Plasma Insulin	-0.01 (0.01)	0.60	2.48	1.00
		BMI-adjusted Fasting Plasma Insulin	-0.01 (0.01)	0.22	1.84	1.00
		Corrected Insulin Response Adjusted for Insulin Sensitivity	0.03 (0.14)	0.84	9.01	0.91
		Plasma Glucose at 120min during OGTT	-0.24 (0.21)	0.27	6.79	0.98
		Fasting ProInsulin	0.03 (0.05)	0.47	5.07	1.00
		Overall Insulin Response during OGTT	-0.07 (0.14)	0.64	15.53	0.56

Extra columns are shown on the next page

MR Egger regression method			Median weighted MR method		Penalised median weighted MR method	
Beta (SE) for continous variable OR [95% CI] for binary variable	*p*-value	Egger Intercept term (P value)	Beta (SE) for continous variable OR [95% CI] for binary variable	*p*-value	Beta (SE) for continous variable OR [95% CI] for binary variable	*p*-value
0.70 [0.58-0.85]	2.42E-04	0.24	0.88 [0.77-1.00]	0.05	0.92 [0.80-1.05]	0.22
0.82 [0.66-1.01]	0.07	0.15	0.97 [0.86-1.10]	0.61	0.96 [0.85-1.09]	0.56
0 (0.03)	0.86	0.55	0.01 (0.02)	0.73	0.01 (0.02)	0.63
0 (0.03)	0.89	0.27	0.04 (0.02)	0.01	0.04 (0.02)	0.01
0.10 (0.02)	5.19E-05	0.09	-0.01 (0.02)	0.78	-0.03 (0.02)	0.08
0.12 (0.03)	6.84E-06	0.06	-0.02 (0.02)	0.21	-0.03 (0.02)	0.11
-0.01 (0.03)	0.83	0.68	-0.01 (0.02)	0.46	-0.02 (0.02)	0.44
0.13 (0.07)	0.04	0.47	0.1 (0.04)	0.01	0.1 (0.04)	0.01
0.01 (0.07)	0.94	0.23	0.04 (0.04)	0.35	0.04 (0.05)	0.38
-0.03 (0.07)	0.68	0.14	0.07 (0.04)	0.09	0.07 (0.04)	0.11
-0.09 (0.07)	0.19	0.51	-0.07 (0.04)	0.12	-0.08 (0.04)	0.07
0.14 (0.03)	8.50E-07	0.23	0.06 (0.02)	7.42E-04	0.06 (0.02)	9.72E-04
0.16 (0.03)	2.73E-08	0.51	0.11 (0.02)	1.75E-08	0.11 (0.02)	2.16E-08
0.2 (0.16)	0.20	0.16	0.06 (0.09)	0.53	0.06 (0.09)	0.53
-0.53 (0.34)	0.12	0.06	-0.01 (0.19)	0.94	-0.01 (0.18)	0.94
-0.02 (0.05)	0.66	0.24	0 (0.03)	1.00	0 (0.03)	0.97
-0.03 (0.03)	0.43	0.17	0.02 (0.02)	0.40	0.02 (0.02)	0.32
-0.05 (0.03)	0.16	0.06	0.01 (0.02)	0.72	0.01 (0.02)	0.66
-0.04 (0.03)	0.24	0.47	-0.01 (0.02)	0.48	-0.02 (0.02)	0.24
-0.05 (0.03)	0.07	0.29	-0.02 (0.01)	0.12	-0.03 (0.01)	0.06
-0.46 (0.35)	0.18	0.05	-0.18 (0.2)	0.37	-0.18 (0.19)	0.34
-0.25 (0.53)	0.64	0.98	-0.1 (0.28)	0.72	-0.1 (0.29)	0.73
0.12 (0.12)	0.31	0.45	0.06 (0.07)	0.40	0.06 (0.07)	0.38
-0.72 (0.35)	0.04	0.03	-0.13 (0.2)	0.51	-0.13 (0.2)	0.50

B

Independent variable	Dependent variable		Inverse variance weighted MR method		Heterogeneity test	
			Beta (SE) for continous variable	*p*-value	Cochran's Q test	
	Category	Phenotype	OR [95% CI] for binary variable		Cochran's Q	*p*-value
One-SD increase of genetically predicted LTL	Disease	CHD	1.05 [0.99-1.10]	0.08	24.34	1.00
		T2D	1.06 [1.01-1.08]	0.01	27.21	1.00
	Adiposity	BMI	0.01(0.01)	0.09	14.76	1.00
		BMI-adjusted Wasit/Hip Ratio	0(0.01)	0.83	18.51	1.00
		BMI-adjusted Wasit Circumference	0(0.01)	0.70	19.11	1.00
		BMI-adjusted Hip Circumference	0(0.01)	0.70	19.11	1.00
		Body Fat Percentage	0.03(0.01)	4.51E-03	21.11	1.00
	Blood Pressure	Diastolic BP	-0.01(0.01)	0.25	18.17	1.00
		Systolic BP	-0.01(0.01)	0.33	19.08	1.00

Extra columns are shown below.

MR Egger regression method			Median weighted MR method		Penalised median weighted MR method	
Beta (SE) for continous variable	*p*-value	Egger	Beta (SE) for continous variable	*p*-value	Beta (SE) for continous variable	*p*-value
OR [95% CI] for binary variable		Intercept term	OR [95% CI] for binary variable		OR [95% CI] for binary variable	
0.99 [0.87-1.1]	0.90	0.52	1.05 [0.94-1.14]	0.37	1.05 [0.94-1.14]	0.36
1.08 [0.97-1.12]	0.17	0.86	1.01 [0.94-1.04]	0.74	1.01 [0.93-1.04]	0.85
-0.03(0.02)	0.20	0.23	-0.02(0.02)	0.28	-0.02(0.02)	0.30
0(0.02)	0.86	0.97	-0.01(0.02)	0.42	-0.02(0.02)	0.16
-0.04(0.02)	0.05	0.40	0.01(0.02)	0.69	0.01(0.02)	0.49
-0.04(0.02)	0.05	0.40	0.01(0.02)	0.69	0.01(0.02)	0.49
-0.05(0.02)	0.02	0.07	0.01(0.02)	0.58	0.02(0.02)	0.27
-0.09(0.02)	2.88E-05	0.06	-0.03(0.02)	0.09	-0.03(0.02)	0.07
-0.08(0.02)	2.36E-04	0.13	-0.02(0.02)	0.20	-0.02(0.02)	0.34

3.3.3 Genetic correlations to a variety of human phenotypes and diseases

We explored human diseases and traits that shared common genetic aetiologies with LTL by performing LDSC analyses that tested genome-wide genetic correlations between LTL and 320 selected traits and diseases curated within the LD hub[300,301] (section 3.2.3). In comparison to the MR approach, these analyses utilise overall GWAS results rather than selected SNPs with the most significance. In agreement with our MR analyses, LTL was negatively genetically correlated with CAD (r=-0.17, *p*-value=0.01, Supplementary Table 12). In contrast to the MR results in the section 3.3.2.3, genetic correlations of LTL with dyslipidaemic risk factors were all in the negative direction, i.e. longer LTL genetically correlated with lower levels of these risk factors, therefore directionally concordant with the correlation with CAD. These risk factors included LDL and total cholesterol, total triglycerides and HDL cholesterol levels (Supplementary Table 12). These suggested shared genetic architecture underlying LTL, CAD and CAD risk factors. However, it should be noted that despite some correlations observed, the levels of significance were nominal and did not reach the Bonferroni corrected threshold.

3.3.4 PheWAS in UK Biobank

Telomere homeostasis is important for suppressing tumorigenesis and metastatic malignant transformation[313,314]. LTL has previously been associated with risks of overall and site-specific cancers[158,177,208,315], but causations remain controversial, and significant findings so far have been restricted to certain types of diseases that had large-scale GWAS results available. Large homogenous cohorts with a more comprehensive coverage of a variety of diseases can help to identify associations with additional diseases that have not been studied before, and provide additional reference for uncertain causations, yet only when sufficient cases are present in such cohorts, which may include cardiometabolic disorders and some common cancers. To refine causal associations of LTL with diseases previously reported and discover novel associations of LTL with a broader range of diseases, I used a dual approach: PheWAS in a smaller but manually refined subset of clinical outcomes, and a larger yet crude full set of ICD10-codes defined clinical outcomes in UK Biobank.

3.3.4.1 Manually refined subset of clinical outcomes

Investigating 122 curated outcomes in UKBB, a total of 30 nominally significant associations were identified, nine of which passed the Bonferroni corrected threshold (p-value<$4.1x10^{-4}$, Figure 3.3, Supplementary Table 15). These included novel findings of decreased risk of hypothyroidism, and increased risks of thyroid cancer, lymphoma and diseases of excessive growth (uterine fibroid, uterine polyps and benign prostatic hyperplasia) for individuals with longer LTL. Moreover, in line with previous findings genetically predicted longer LTL was associated with decreased risk of CAD (p-value=0.01) and significantly increased risks of lung and skin cancers and leukaemia after multiple testing correction[158,177,185,186,285,291,313,314,316–321]. Our results also supported a causal role of longer LTL in reducing risks of rheumatoid arthritis, aortic valve stenosis, chronic obstructive pulmonary disease (COPD) and heart failure, all of which have previously been associated with LTL in prospective, retrospective and MR studies[158,178,179].

3.3.4.2 Full set of ICD10-codes defined clinical outcomes

In contrast to previously published findings that highlighted carcinogenic sites with lower rates of stem cell division to be more susceptible to genetic differences in LTL[158], we found various tissues with higher proliferative capacity to have strong associations with LTL, including haematological malignancies, male genital and prostate cancer, melanoma and malignant neoplasms in epidermal tissue. Similarly, benign neoplasms and non-neoplastic disorders were also found more likely to be associated in such tissues (Figure 3.4, Supplementary Table 17). It seemed that significant diseases (p-value<0.05) were concentrated in two clusters, neoplasms and diseases in genito-urinary system (Figure 3.5).

Over one third of the strongly associated diseases exhibited high levels of between-variants heterogeneities (Cochran's Q test p-value <0.05, Figure 3.4, Supplementary Table 17). Some showed locus-specific effects, for example, intestinal malabsorption was heavily driven by the *HLA* locus, excluding which decreased the estimated causal effect by more than 70% (OR[95%CI]=0.95[0.92-0.97], Figure 3.4, Supplementary Table 17). Other results did not change much after excluding the *HLA* locus. Moreover, malignant neoplasm of the brain, of which 80% cases were attributed to glioma[322], one of the most significantly associated cancers found by previous studies[158,323], was specifically associated with the *TERT* locus (OR[95% CI] per risk allele=1.50[1.27-1.78], p-value =$2.05X10^{-6}$), but not genetically determined LTL using

all genome-wide independent variants combined. Some specific diseases showing evidence of associations with individual LTL loci were phenotypically reminiscent of monogenic telomere syndromes, for instance, seborrheic keratosis was associated with *TERC* and *TERT* loci and reminiscent of dyskeratosis congenita, as both of which are commonly featured with dermatological dystrophy.

Figure 3.3. MR results for effects of shorter LTL on risks of 122 diseases in UK Biobank.
Data shown are ORs and 95% CIs per 1-SD shorter genetically predicted LTL. LTL is genetically predicted using independent variants with FDR<0.05. Diseases are classified into groups as indicated by boxes and sorted alphabetically within each disease group. Nominally significant (*p*-value<0.05) causal associations estimated via inverse-variance weighted MR method are shown in green for a reduction in risk and red for an increase in risk due to shorter LTL. Where ^{O} indicates nominal (*p*-value<0.05) evidence of pleiotropy estimated by MR-Egger intercept. Full results are also shown in Supplementary Table 15 along with full MR sensitivity analyses.

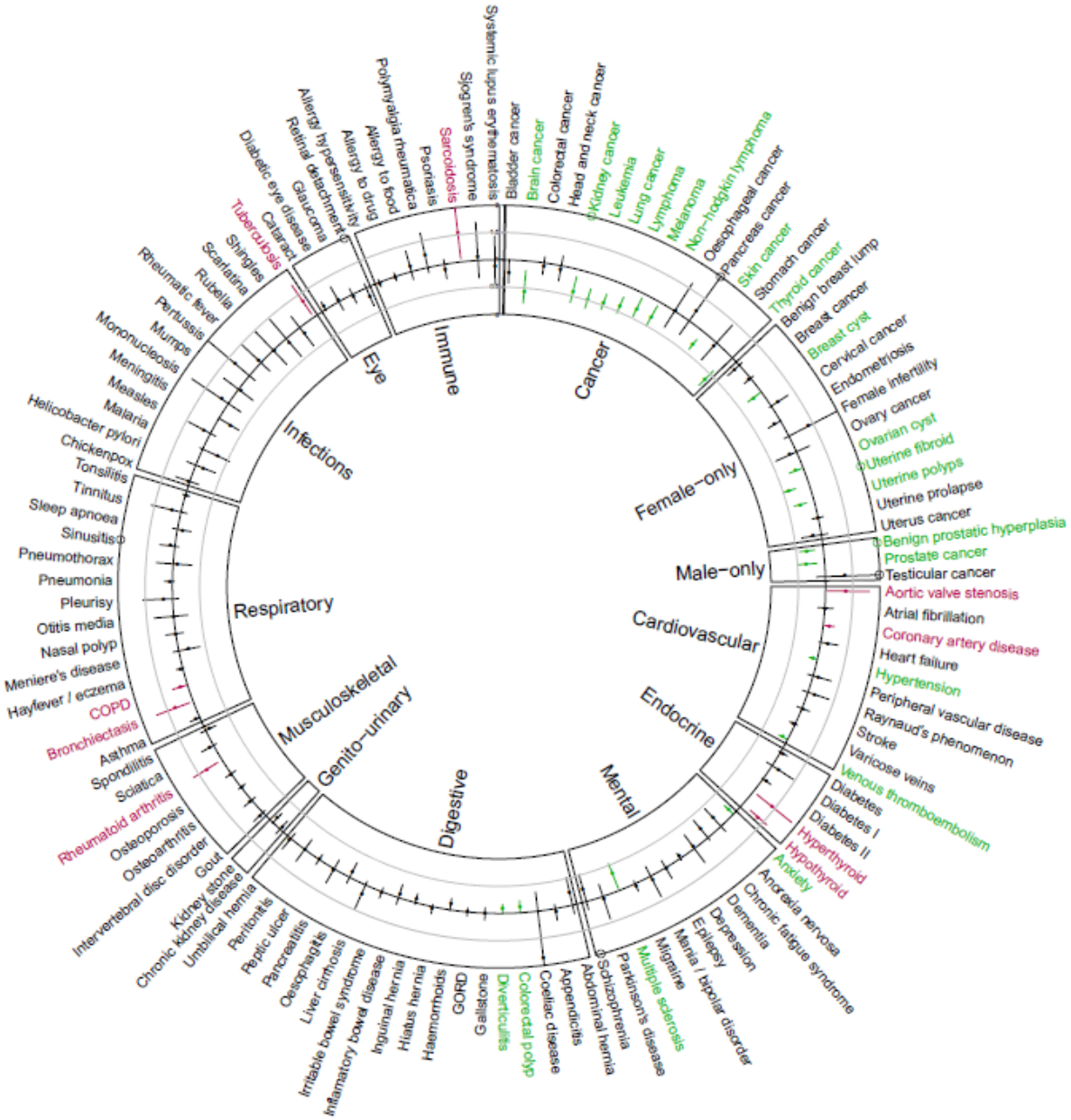


Figure 3.4 Significantly associated diseases with longer LTL estimated using genome-wide significant independent lead variants.
The forest plot shows significant associations with genetically determined LTL, and the heatmap shows z-scores of individual variant associations with these diseases.

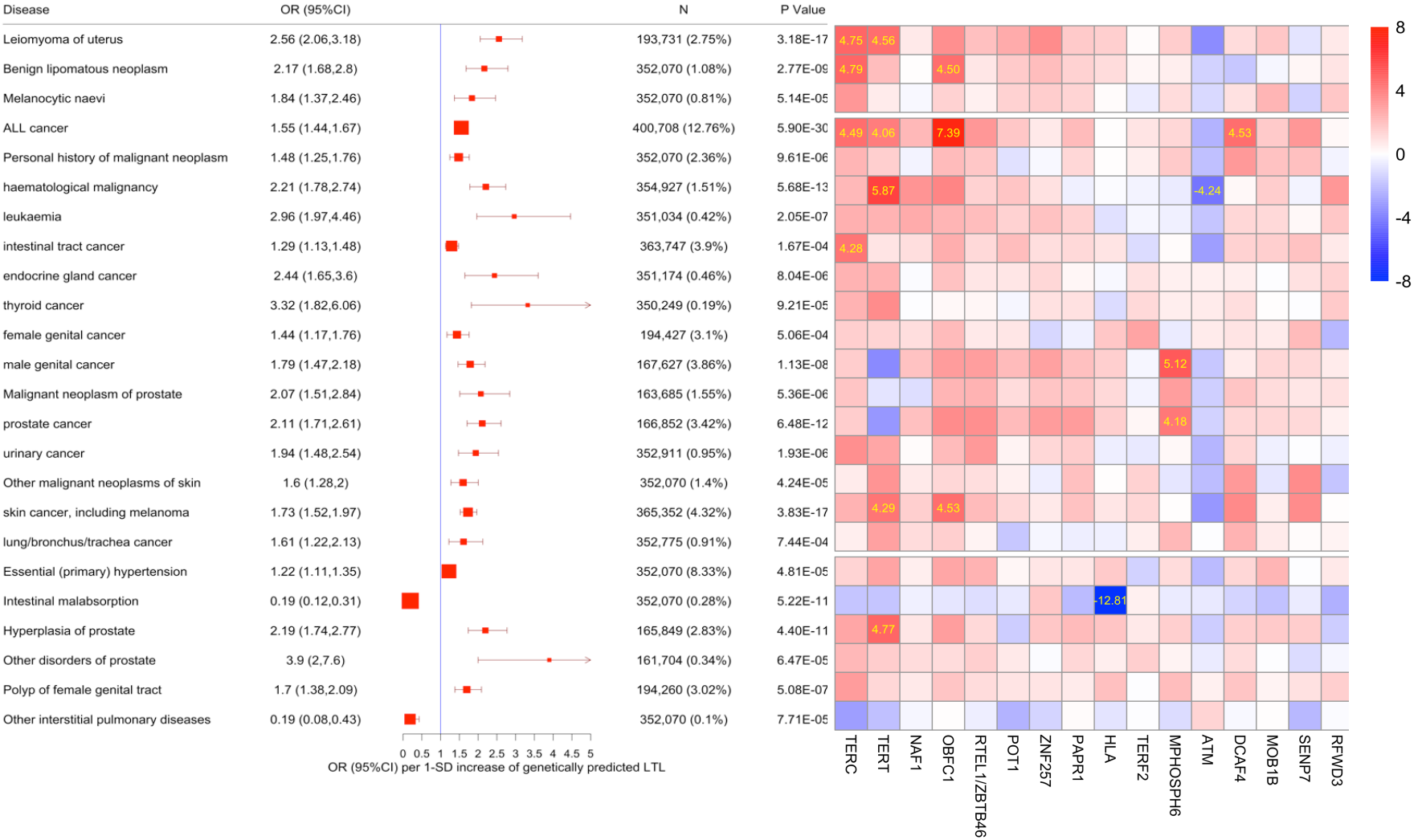
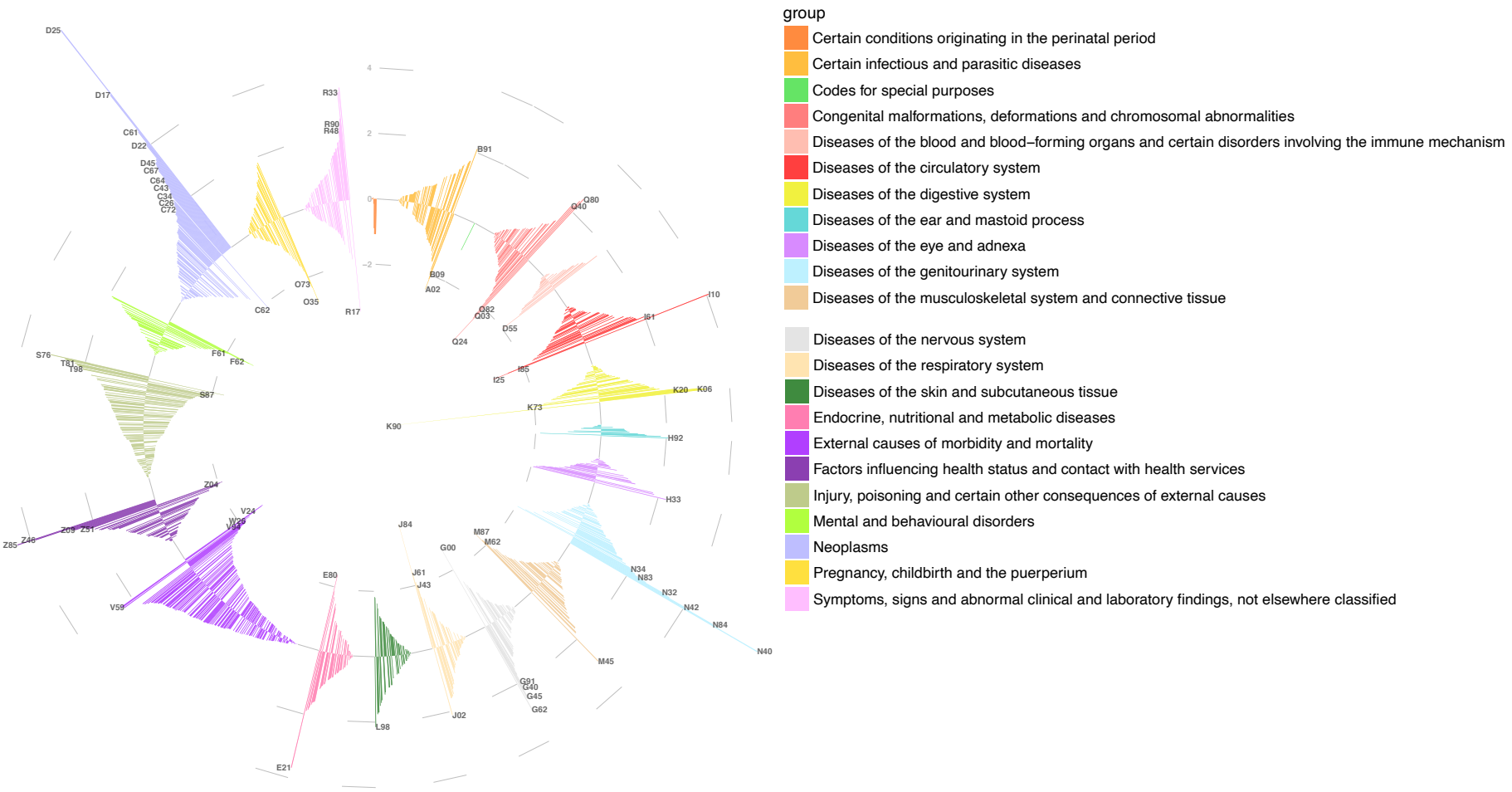


Figure 3.5 Circular plot of PheWAS of LTL.

LTL is estimated using independent lead variants at genome-wide significant loci. All ICD10-coded disease outcomes are clustered into 22 categories colour labelled. Z-scores were plotted as bar heights. Diseases with absolute z-scores larger than 1.96, corresponding to a nominal significance level, were labelled with ICD10 codes besides bars.



3.3.5 Single-locus based cross-phenotype associations

We also examined individual locus-driven genetic correlations between LTL and a variety of human phenotypes and diseases by querying 52 conditionally independent variants at FDR loci and their closely-related SNPs in LD ($r^2 \geqq 0.8$) across publicly available GWAS databases using PhenoScanner[312]. While some morbidities showed specific correlations to individual loci, others were correlated to a broader spectrum of loci. For example, self-reported hypothyroidism or myxoedema exhibited strong associations particularly with the *TERT* locus, which was also exclusively responsible for several subtypes of ovarian cancers. Moreover, *Leucine-Rich Repeat-Containing Protein 16A* (*LRRC16A*) gene located near the *HLA* region was responsible for several types of gastrointestinal disorders, including intestinal malabsorption, coeliac disease and primary sclerosing cholangitis. In contrast, some diseases or traits were associated with multiple LTL loci, including blood cell traits and haematological diseases that involved *TERC*, *TERT*, *LRRC16A*, *SENP7*, *ATM*, SAMHD1, and ENTPD5 (*Ectonucleoside Triphosphate Diphosphohydrolase 5*); similarly, respiratory function and lung cancers involved *TERC*, *TERT*, *LRRC16A*, *OBFC1* and *MPHOSPH6*, suggesting multiple common genetic attributes that were shared between LTL and these phenotypes and diseases.

3.4 Discussion

This chapter substantially expands our current knowledge on potential impacts of telomere dysregulation on cardio-metabolic traits, cancers and a broad spectrum of human diseases, providing insights onto roles of LTL in disease susceptibilities, and how variation of individual LTL-associated genes affects disease risks.

3.4.1 Clinical relevance of genetically predicted LTL

We have not only confirmed previous findings linking genetically predicted shorter LTL to a higher risk of CAD and lower risks of several cancers, but also demonstrated novel causalities of LTL on thyroid diseases and cancers, lymphoma and several non-malignant neoplasms. Notably, shorter genetically predicted LTL was found to be protective for all of these

proliferative disorders, potentially through limiting cell proliferative capacity, which in turn reduced occurrence of potential oncogenic mutations that can occur during DNA replication. Furthermore, we also provided evidence showing genetic predisposition to shorter LTL increased risks of several cardiovascular, inflammatory and respiratory disorders that have previously been linked to LTL in observational epidemiological studies[158,173].

3.4.2 Association of LTL with cardio-metabolic disorders

No observational evidence has been found to support an association between LTL and incident T2D. Primary MR analysis using independent genome-wide significant variants as genetic instruments have also suggested no evidence for an association between LTL and T2D risk, which was further supported in sensitivity analyses by excluding the *HLA* region, or using variants at FDR<0.05 as instruments or different MR approaches.

The association between genetically predicted longer LTL and decreased risk of CHD was only observed when using genome-wide significant but not FDR variants, with substantial levels of between-variants heterogeneity in both analyses, suggesting several specific loci with larger effect sizes and more precise estimates drive the overall protective effects of longer LTL on CHD.

Conventional risk factors for CHD all showed significant but positive associations with genetically predicted longer LTL, implying that these factors do not mediate the protective effect of longer LTL on CHD, or the MR analyses are not powerful enough to capture true causal associations. These included higher triglyceride, total and LDL cholesterol levels, higher BMI, waist circumference and body fat percentage and blood pressure. The associations with these risk factors could be driven by specific loci that were distinct from those for CHD. On the contrary, genome-wide genetic correlation analyses demonstrated directionally concordant correlations of LTL to CAD and CAD-related risk factors that included lipid profiles of dyslipidaemia, adiposity-related traits and smoking. The discrepant results between MR and genome-wide genetic correlation analyses on the CVD-related risk factors might be due to different genetic factors employed in the two approaches, with the former completely driven by a limited number of significant loci from GWAS, whereas the latter by millions of variants covering the whole-genome. Hence the MR results might be heavily driven by specific loci that showed relatively large causal effects on the CVD-related risk factors, whereas the

genetic correlation results reflected an overall consistency of associations with LTL and CVD-related risk factors at genome-wide scale. Further investigation might be needed to study specific driving forces that led to the directional discrimination of causalities of LTL on CHD and CHD-related risk factors.

Chapter 4

Feasibility of studying longitudinal change of LTL

Abstract

Background Causes and consequences of prospective longitudinal changes of LTL may differ from those of LTL measured once at study baseline. Few studies have attempted to analyse the differences, and little consensus has been reached.

Objectives To conduct a systematic literature review on the longitudinal change of LTL and evaluate feasibility of studying it in a relatively young and healthy prospective cohort in the Fenland study.

Methods A systematic review was conducted for repeated measures of LTL over time. Searches were performed in PubMed of studies published from January 2009 to January 2018 using the following search strategy: telomeres AND (shortening OR lengthening) AND (cohort studies OR genetics). Quality of studies was assessed, and results were extracted. A pilot analysis was performed within the Fenland study cohort, where two distinct groups of individuals were selected on the basis of different durations of follow-up time: one (n=14) with 3-5-year, and the other (n=40) with 8-10-year intervals.

Results Sixty-five papers were included in the final set of eligible studies, and these showed differences in terms of study designs, demographics of participants, methods of LTL measurement and statistical approaches. Reported results were inconsistent, except for baseline LTL as a strong determinant of longitudinal changing rate of LTL, which was also observed in the Fenland pilot study. Annual changing rates were comparable between the shorter (3-5-year) and longer (8-10-year) time intervals studied.

Conclusions Changes in LTL are detectable in relatively young and healthy individuals, but are technically challenging and unlikely to be scientifically useful as an outcome or cost-efficient for genetic association studies.

4.1 Introduction

Most previous epidemiological studies of LTL measured LTL at one time point, and thus were unable to investigate longitudinal changes of LTL, for which repetitive measures at more than one time point are required. The few studies that analysed changes over time showed inconsistent results in associations with risk factors and disease outcomes yet no systematic literature review has been conducted. Moreover, studies that compared the results for studying LTL at one *versus* multiple time points (i.e. longitudinal changes of LTL over time), have demonstrated discrepant findings in associations with risk factors and diseases[324,325], emphasising the importance of studying the longitudinal change as a separate trait from one-time measure of LTL. Longitudinal changes of LTL provide additional temporal information in LTL dynamics compared to a snapshot one-time measurement, and may differ from the one-time measurement of LTL in terms of their associated genetic and non-genetic risk factors, pathological links to diseases and potential suitability as a biomarker of LTL-associated diseases.

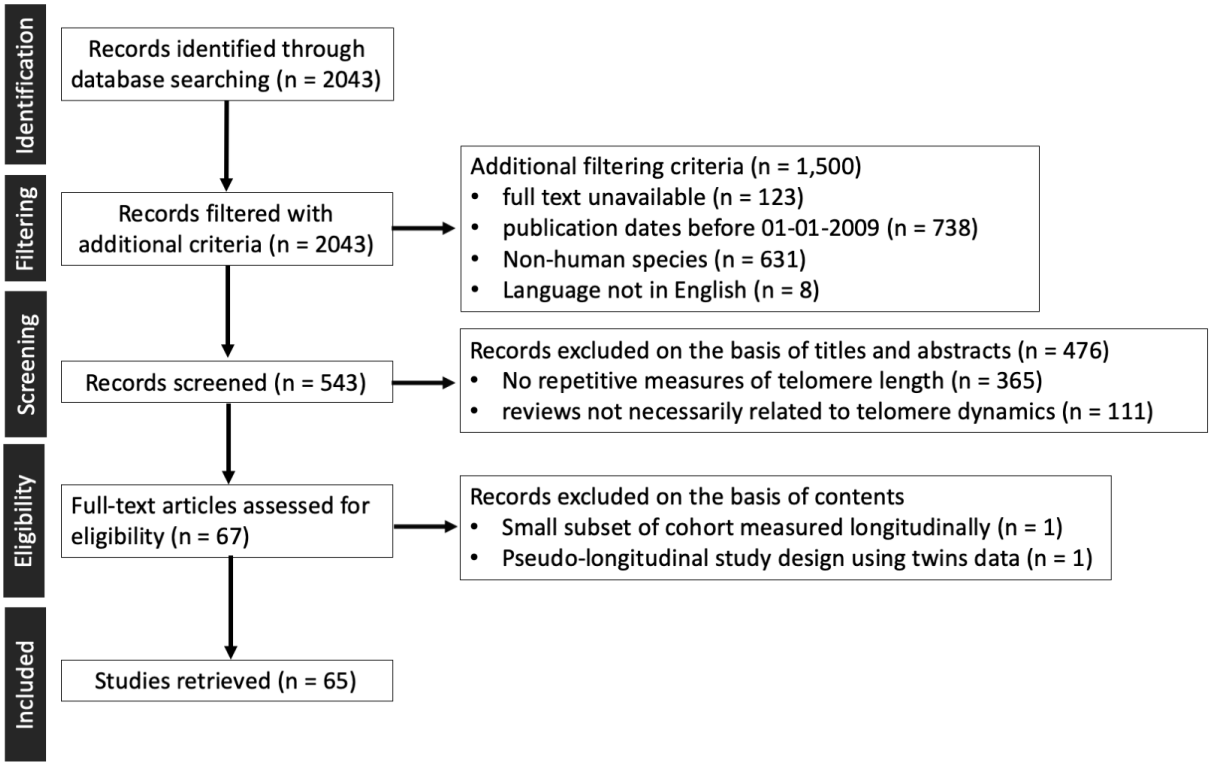
Despite the clear rationale and scientific importance of studying longitudinal changes of LTL, there are many technical challenges, such as sensitivity and accuracy of the LTL measurements and statistical issues arising when repeated measures are made on the same individuals, such as the regression to the mean problem that makes true longitudinal changes difficult to distinguish from random fluctuation of LTL measures. Therefore, I conducted a feasibility test by first systematically reviewing epidemiological studies of the longitudinal changes of LTL, and then initiated a preliminary test in a population-based prospective cohort study in Fenland, where LTL was measured at two time points in 40 and 14 individuals with 8-10-year and 3-5-year time intervals, respectively, to assess influences of different duration time on longitudinal measures. In addition, seventy-four individuals were measured at baseline and evenly clustered into 10-year age bins to confirm a correlation of baseline LTL with age.

4.2 Methods

4.2.1 Systematic Literature Review

A systematic literature review was conducted following the MOOSE guideline[326], in order to critically evaluate and summarise previously published research on longitudinal changes of LTL. The search strategy, applied to the PubMed database, used a combination of terms or their synonyms: “telomeres” AND (“shortening” OR “lengthening”) AND (“cohort studies” OR “genetics”), with details provided in the Supplementary Notes. Records without full text available online, published prior to 2009 when the real time qPCR method for LTL measurement was published[327], studies in species other than humans and those not written in English were excluded from the search prior to screening titles and abstracts. When reviewing titles and abstracts, exclusion criteria included studies without repetitive measures of LTL or reviews not related to telomere dynamics. The articles that passed the filtering criteria were reviewed for their full text and excluded if the study design was not longitudinal or the longitudinal component only involved a small subset of the study, resulting in a total of 65 papers included into the final set of eligible studies (Figure 4.1).

Figure 4.1: Flow-chart of the systematic literature search for epidemiological studies of longitudinal telomere changes.



4.2.2 LTL changes over time in the Fenland study

4.2.2.1 Study participants and design

The Fenland study is a prospective cohort study of 12,435 participants born between 1950 and 1975[328,329]. Between 2005 and 2015 (phase 1), participants were recruited from general practices in Cambridge, Ely and Wisbech (UK), and predominantly healthy. Individuals who were pregnant, previously diagnosed with diabetes, unable to walk unaidedly, or had psychosis or terminal illness were excluded. Metabolic phenotypes and genome-wide genotypes were measured in detail, and the second follow-up of the cohort (phase 2) is ongoing, collecting longitudinal data of the same cohort of participants on key risk factors and continuous metabolic traits. All study procedures were approved by the Health Research Authority National Research Ethics Service Committee East of England-Cambridge Central, and all participants provided written informed consent.

To evaluate feasibility of studying differences in LTL changing rates between different follow-up time intervals, two distinct groups of participants were selected on the basis of different durations between Fenland phase 1 and 2 visits: one group (n=14) with a time interval of 3-5 years, and the other (n=40) with 8-10 years. In addition, there were 74 samples measured at phase 1 only and clustered into several age groups (25-35, 35-45, 45-55 and 55-60 years), which were used to confirm correlations of baseline LTL and age.

4.2.2.2 Sample preparation and DNA extraction

All blood samples were collected in tubes of ethylenediaminetetraacetic acid using standard venepuncture protocols, with plasma, serum and buffy coat aliquots extracted by differential velocity centrifugation and stored in -80 °C prior to use[330]. Following a standard protocol, DNA was extracted via the Autopure method (Qiagen) for phase 1 plasma samples, and for phase 2 samples, the Promega ReliaPrepTM Large volumeHT gDNA isolation system (Promega, A2751) was used, coupled with a Tecan EVO automated liquid handling platform with integrated HSM 2.0 Heater shaker Magnet (Promega). DNA samples were diluted 25-folds to 2 ng/µL, and 5 µL of the diluted DNA samples were used in all experiments.

Control DNA samples were used for each plate, which consisted of negative and positive controls and standard curves. Negative control: nuclease-free water; positive control: pooled phase 1 samples that contained 12 participants, two for each gender at age 20, 40 and 60

years, and genomic DNA (G304A, Promega) at low (1.56ng), medium (6.25ng) and high (25ng) absolute concentrations; standard curves: genomic DNA (G304A, Promega) at concentrations of 50ng, 25ng, 12.5ng, 6.25ng, 3.13ng and 1.56ng, and genomic DNA (K562, Promega) at concentrations of 50ng, 25ng, 12.5ng, 6.25ng, 3.13ng and 1.56ng.

4.2.2.3 LTL measurements

LTLs were measured in DNA samples extracted from blood in both Fenland phase 1 and 2 samples, using the ViiA™ Real-Time PCR System with monochrome multiplex qPCR method[331]. The telomeric DNA was amplified simultaneously with a single copy housekeeping gene, the *albumin* gene. The primer sequences were designed using the validated method described in detail by Cawthon *et al.*[331]. LTL was calculated as the T/S ratio, in which T represents the amount of standard DNA (ng) that matches the experimental sample for copy number of the telomere template, divided by S, the amount of standard DNA (ng) that matches the experimental sample for copy number of the *albumin* gene. Each experimental sample was assayed in triplicate, and the mean of three replicative measures was reported as the final LTL estimate for each sample, following the standard protocol[331].

4.2.2.3.1 Initial experimental set-up

We used two real time qPCR reagents: the GoTaq™ qPCR Master Mix (Promega) and the SYBR Select qPCR Mastermix (Thermo), in combination with two standard curves: K562 and gDNA standard curves. In order to compare experimental performance of different combination of these settings, we used 5 individuals in the initial test, that were randomly selected among Fenland participants with DNA samples available at two phase visits. Experimental reaction and program settings were shown below (Table 4.1). LTL of each individual was measured using DNA samples collected at each phase visit, together with control samples, all analysed in triplicate. DNA samples were dispensed before adding reaction mix. We compared experimental efficiencies of standard curve samples against the optimum efficiency parameters recommended from manufacture guidelines (R^2>0.99, slope coefficients = [-3.6, -3.1], Efficiency(%) = [90,110]).

The qPCR protocols were optimised based on experimental efficiencies using different reagents. The protocol that used the combination of genomic DNA standard curve and GoTaq

reagent exhibited the best performance, and it was the only one that satisfied the optimum efficiency threshold recommended by the manufacture guideline (Table 4.2). The correlations between age and LTL measures were stronger when using the best-performance experimental data, which also exhibited larger intraindividual differences of LTL measures between two time points (Figure 4.2).

Table 4.1 Initial qPCR experimental reaction and program settings:

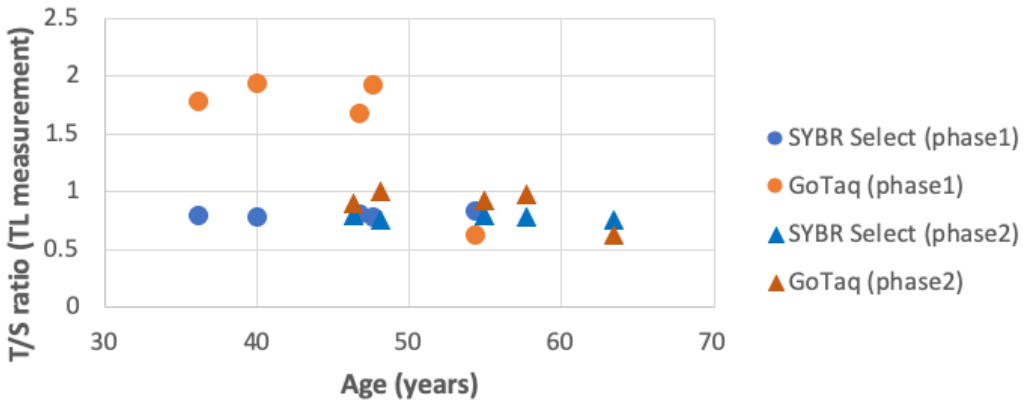
Reaction setting per well	
Total reaction volume	15ul
DNA samples (2ng/ul)	5ul
SYBR Select or GoTaq Mastermix	7.5ul
Forward Primer (10mM)	0.3ul + 0.3ul
Reverse Primer (10mM)	0.3ul + 0.3ul
Nuclease free water	1.3ul
Programme setting per 96-well plate	
Primer concentration (nM)	300
Mastermix (lot number)	SYBR Select (1705053) and GoTaq (0000280334)
Pipetting method	Repeater 1.0mL
Singleplex or duplex	Duplex
Hold stage 50°C 2min	Yes
Duration of hot start (minutes)	2
Annealing temperature (Tel)	62
Annealing temperature 1 (Alb)	84
Number of cycles	32
Signal acquisition temperature (Tel)	73
Signal acquisition temperature (Alb)	87

Table 4.2 Efficiency parameters of qPCR experiments for the two standard curves with different reagents (N=5).
The two standard curves were generated using standard gDNA or DNA extracted from K562 cell lines. Reaction reagents were SYBR Select or GoTaq.

K562 standard curve				
	SYBR Select		**GoTaq**	
	Telomere	***Albumin***	***Telomere***	***Albumin***
R^2	0.999	0.998	0.998	0.997
Slope	-3.798	-3.788	-4.183	-3.892
Efficiency (%)	83.354	83.650	73.412	80.683
gDNA standard curve				
R^2	0.989	0.997	0.997	0.996

Slope	-4.172	-3.625	-3.588	-3.604
Efficiency (%)	73.655	88.729	89.993	89.439

Figure 4.2 Correlations between age and LTL measures for protocols using two different reagents (SYBR Select and GoTaq, coded with different colours). Phase 1 and phase 2 measures for each of the 5 individuals were plotted according to their ages at phase 1 and 2, respectively.



4.2.2.3.2 Scaling-up of the experimental set-up

To examine whether the qPCR protocol can maintain a satisfactory performance when undertaken with larger sample sizes, we expanded sample sizes in three settings as described in the study design: (A) 40 individuals with 8-10-year time intervals, (B) 14 individuals with 3-5-year time intervals, (C) 74 samples at phase 1 (baseline) only. Half of the samples in setting A (20 samples) were measured twice on different plates in order to assess plate effects. Correlation of inter-plate measurement was high (Pearson's correlation coefficient r^2=0.95), suggesting no between-plate heterogeneity. The same best-performance reaction system was applied as in the smaller-scale set-up for all settings (Table 4.4). For reaction program settings, all steps stayed the same, except that the 2-minute hold stage was removed. Each experimental unit (one sample at each time point or a control sample) was tested in triplicate.

The setting A failed to produce a satisfactory performance, i.e. efficiency parameters from this experimental setting were below recommended standards (Table 4.3). Technical issues that might contribute to the lack of efficiency were investigated with multicomponent plots (fluorescence vs. cycle). These plots were spontaneously generated for every single reaction unit, serving as a technical surveillance for the real-time qPCR instrument. We noted that in certain regions of the plate, specifically the upper half (row A-G), and not the lower

half (row I-P) of the plate, amplification curves exhibited abnormal shapes (Figure 4.2), implying an underperformance of the machine. In the settings B and C, we switched each half of the plates, and with extra care for sample loading and dispersion, efficiencies of both targets (Telomere and Albumin) improved and satisfied the optimum threshold, as shown in Table 4.3.

Table 4.3 qPCR efficiency parameters in different experimental settings.

	gDNA (40 samples, 8-10-year intervals)		***gDNA (14 samples, 3-5-year intervals)***		***gDNA (74 samples only at baseline)***	
	Tel	***Alb***	***Tel***	***Alb***	***Tel***	***Alb***
R^2	0.997	0.997	0.998	0.998	0.995	0.999
Slope	-4.025	-3.717	-3.55	-3.582	-3.282	-3.469
Efficiency (%)	77.188	85.792	91.275	90.201	100.000	94.193

Figure 4.2 Multicomponent plots for the qPCR reactions.
The left panel shows the upper half of an exemplar plate, indicating normal qPCR performance; whereas the right panel shows the lower half the same plate, which contains abnormal reactions marked with irregular curve shapes shown as distortions towards the ends and dips in the middle.

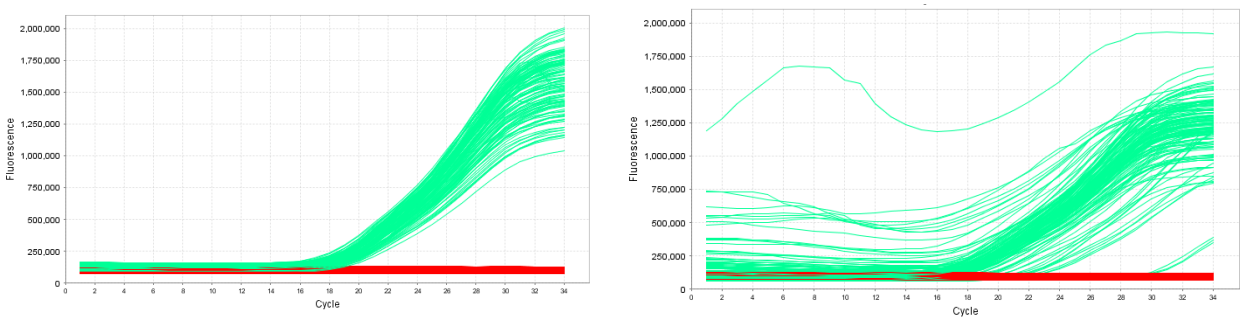


Table 4.4 The final optimised qPCR protocol, including the reaction system and the program setting.

Reaction setting per well	
Total reaction volume	15ul
DNA samples (2ng/ul)	5ul
GoTaq (Promega)	7.5ul
Forward Primer (10mM)	0.3ul + 0.3ul
Reverse Primer (10mM)	0.3ul + 0.3ul
Nuclease free water	1.3ul

Program setting per 96-well plate	
Primer concentration (nM)	300
Standard curve DNA	G304A (Promega)
Mastermix	GoTaq (Promega)
Pipetting method	Repeater 1.0mL
Singleplex or duplex	Duplex
Hold stage 50°C 2min	No
Duration of hot start (minutes)	2
Annealing temperature (Tel)	62
Annealing temperature 1 (Alb)	84
Number of cycles	32
Signal acquisition temperature (Tel)	74
Signal acquisition temperature (Alb)	88

4.2.3 Statistical analyses

Correlations between age and LTL at each phase were calculated using the Pearson's correlation test. Linearity and homoscedasticity assumptions were visually examined by drawing scatter plots (Figure 4.4). Longitudinal changing rates of LTL were estimated by subtracting LTL measures at phase 1 from those at phase 2, and dividing the resultant differences by the time interval. Statistical associations of the longitudinal changing rates with age and baseline LTL measures were tested in linear regression models. Mann-Whitney U test was performed to compare whether the longitudinal changing rates differ between shorter (3-5-year) and longer (8-10-year) time intervals. Analyses were conducted in R version 3.5.1.

4.3 Results

4.3.1 Systematic literature review

4.3.1.1 Main characteristics of the studies

Studies identified in this systematic literature review examined associations between LTL attrition rates and a variety of risk factors and clinical outcomes. Main characteristics of these studies varied, including time intervals between measuring points, measuring techniques, health profiles and ethnicities of study participants, sample types and sizes, and risk factors and clinical outcomes tested (Supplementary Notes). Follow-up times ranged from 4 months

in a randomized pilot bio-behavioural clinical study to 41 years in a multi-ethnic cohort study, with most studies (47/65 = 72.3%) exceeding five years of follow-up time. While most of the studies were at small scale, two of them had relatively large sample sizes, the Prevention of RENal and Vascular ENd stage Disease (PREVEND) study (n=8,074) and the Copenhagen City Heart (CCH) study (n=4,576). Most of the studies (42/65 = 64.6%) used qPCR to measure LTL, while there were 17 studies that also employed the terminal restriction fragment analysis, the gold-standard approach to quantify TL, which is not feasible to do at large scale. Peripheral blood leukocytes were the most frequently used cell/tissue types, a few also used cord blood from new-borns, bone marrow compartments and blood cell subtypes retrieved from sophisticated isolation experiments, which included granulocytes, monocytes and lymphocytes and subtypes of circulating immune cells.

4.3.1.2 Factors associated with accelerated telomere attrition in observational studies

Telomeres have been shown to shorten at different rates in men and women and at different ages in population-based studies. The gender difference occurred in an age-dependent manner after adulthood. Annual attrition rates increased with age and were faster on average in men than in women[164,332]. Studies have shown telomere attrition decelerated in women during menopausal transition, while in men the attrition rate tended to increase with age[333]. Moreover, several studies have consistently suggested the baseline LTL as a powerful predictor of telomere attrition rate[334–336]. They found longer telomeres at baseline were associated with accelerated telomere attrition. This could possibly imply a negative feedback regulation of LTL[336], or it may also be simply due to a statistical problem of the regression to the mean because of unsystematic measurement errors that result in random fluctuation within the measurements.

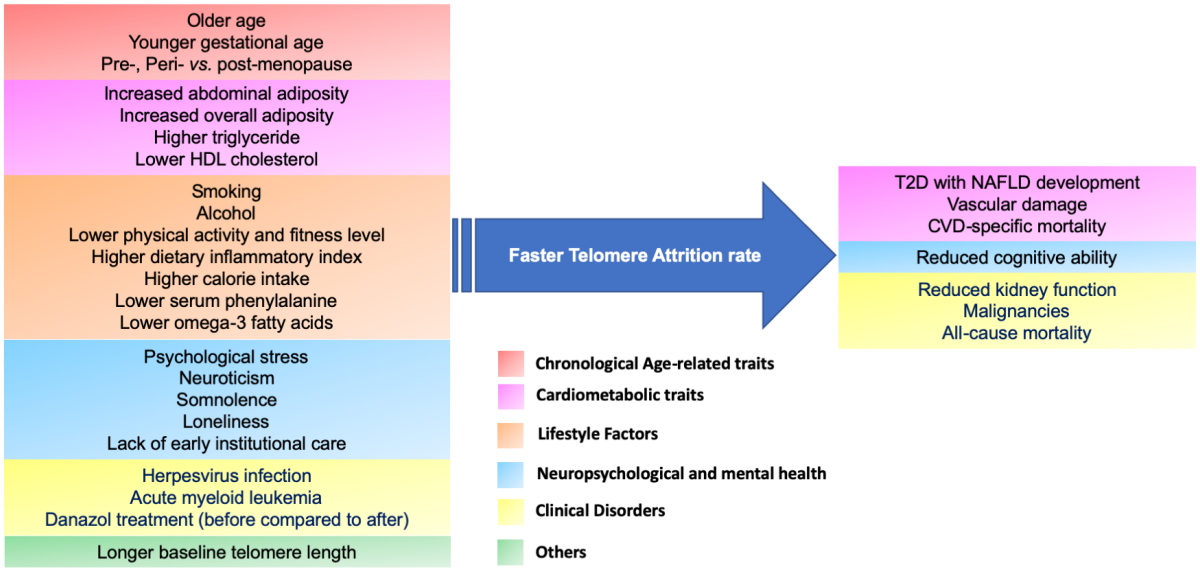
Several lifestyle factors have been identified to accelerate LTL attrition, including smoking[332], alcohol consumption[337], higher energy intake[338], dietary composition (lower levels of serum phenylalanine and omega-3 fatty acids and higher scores of dietary inflammatory index)[339–341], and lower physical activity and fitness levels[342]. Cardio-metabolic traits, such as increased overall or abdominal adiposity, higher triglyceride and glucose levels, lower HDL cholesterol levels[332]; neuropsychological and mental health features, such as psychological stress, neuroticism, somnolence, loneliness and lack of early institutional care[343,344], also increased LTL attrition rates. Moreover, some diseases clinically manifested

with shorter LTL have been reported to increase LTL shortening rates, such as Herpesvirus infection and acute myeloid leukaemia; and on the contrary, the corresponding clinical interventions, such as Danazol treatment, reduce attrition rates[345–347] (Figure 4.2).

However, associations between LTL shortening and factors other than age and baseline TL were inconsistent across studies. For example, in the two largest studies of LTL attrition rates, the PREVEND study[332] and the CCH Study[164], annual attrition rates of LTL were higher in smokers and in participants with greater abdominal obesity in the PREVEND cohort study with 6.6-year of follow-up[332]. In contrast, no associations were found between telomere attrition rate and smoking and body weight in the CCH Study with 10-year of follow-up[164].

Figure 4.3 Overview of determinants and consequences associated with accelerated telomere attrition.

The left block contains changes of the factors that lead to acceleration of telomere shortening, and the right block contains phenotypes that are resulted from faster telomere attrition.



4.3.1.3 Consequences of accelerated telomere attrition

Accelerated telomere attrition, potentially reflecting more rapid impairment of overall genomic integrity along with age, has been associated with higher risks of age-related complex diseases and mortality. In a prospective cohort study of T2D patients, telomeres were found at comparable lengths at baseline and significantly shorter in patients who developed non-alcoholic fatty liver disease than patients who did not during a 6-year follow-up time[348], suggesting that it might be the telomere shortening rather than the baseline TL

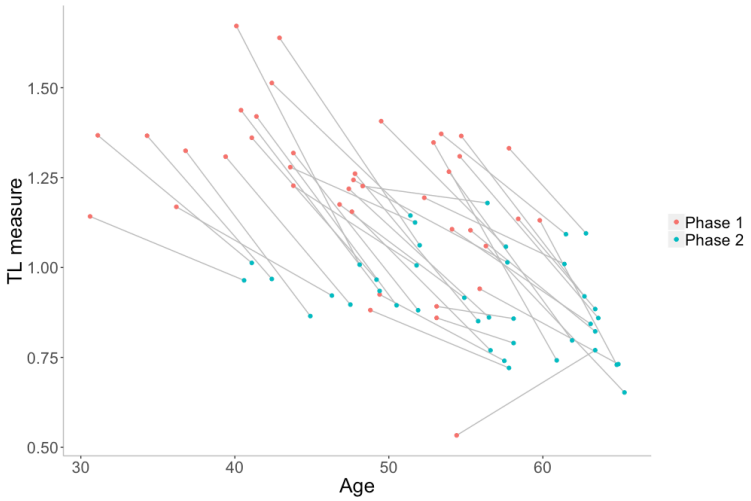
that affected susceptibilities of metabolic disorders. Lending support to this notion, another cohort study on cardiovascular phenotypes also found that individuals with increased rates of telomere shortening, but similar telomere measures at baseline were at higher risks of cardiac and vascular damage[349], as well as cardiovascular mortality[350]. Other diseases have also been associated with telomere attrition rates. For example, patients with haematologic malignancies who carried mutations in genes encoding telomere-associated proteins (e.g. TERT, TERF1, TERF2, ATM and POT1) or showed altered expressions of these proteins have shown accelerated telomere attrition compared to normal volunteers. Accelerated attrition was also seen in patients who received therapeutic treatment of hematopoietic stem cell transplantation[351]. Evidence suggests that genetic variation and clinical intervention can modulate telomere attrition rates, which can in turn influence disease occurrence and progression.

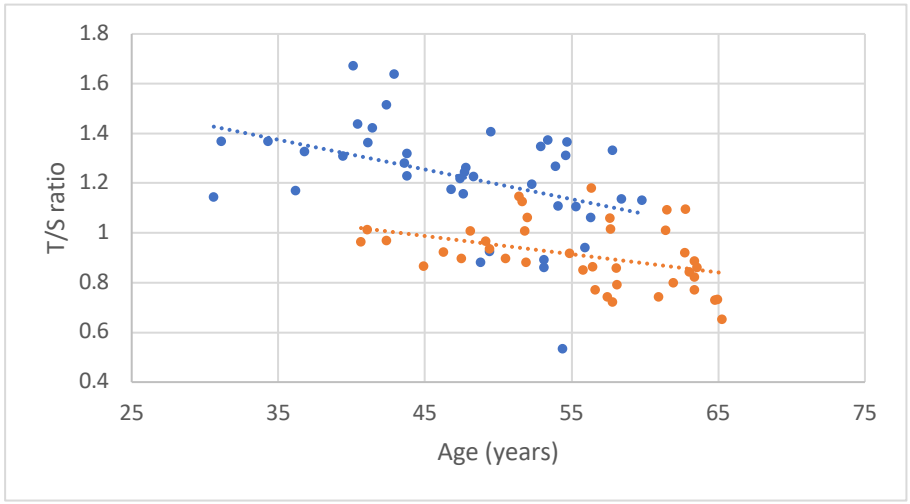
4.3.2 Age correlation to LTL measures at phase 1 and phase 2

For the 40 samples with repeated measurements at 8-10-year intervals, both measures of LTL at phase 1 (mean[SD]=1.22[0.22]) and phase 2 (mean[SD]=0.91[0.13]) were significantly associated with age (*p*-value=0.007 and 0.01, respectively); effect sizes were comparable between the two phases (beta coefficient[SD] = -0.01[0.004] and -0.007[0.003], respectively, Figure 4.4A). The correlation between age and LTL measures was further confirmed within the independent set of 74 samples measured at baseline (beta coefficient[SD]=-0.02[0.004], *P-value*=$5.73X10^{-7}$, Figure 4.4C). For the 14 samples measured at both phases of 3-5-year intervals (mean[SD]=0.99[0.16] and 0.87[0.14] for the phase 1 and 2 measurements, respectively), the correlation with age was not significant in either phase (beta coefficient[SD] = 0.004[0.008], 0.005[0.007], *p*-value=0.63 and 0.47, respectively), possibly due to insufficient power.

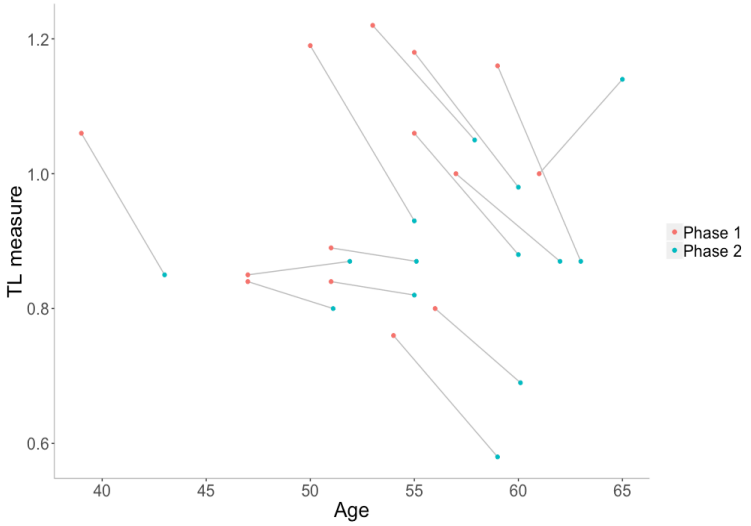
Figure 4.4 Changes of LTL measures at two time points and their associations with age.
On the left panel, LTL measures are plotted against age. Measures at phase 1 are labelled in red, and at phase 2 in green. Two measures of the same individuals are connected. On the right panel, blue and orange dots indicate measures at phase 1 and 2, respectively.

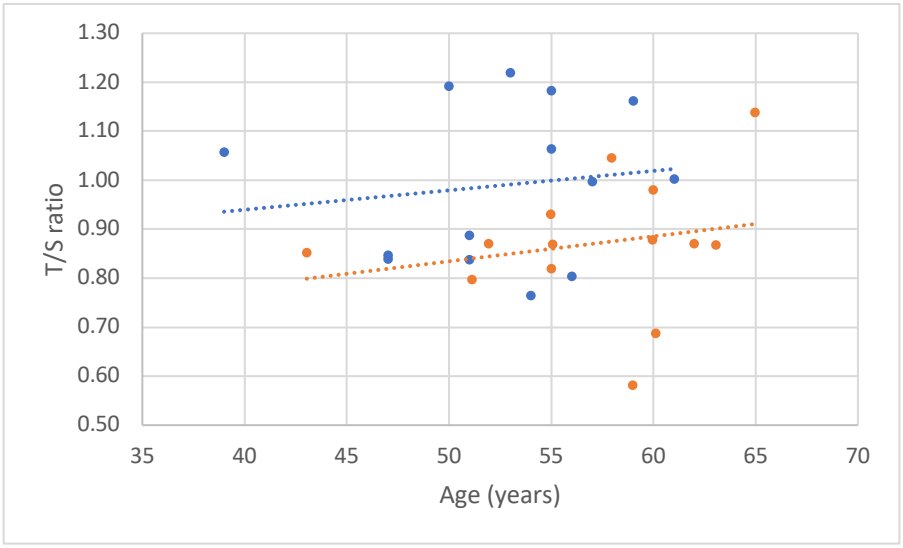
A. Forty samples measured at two phases with 8-10-year intervals.



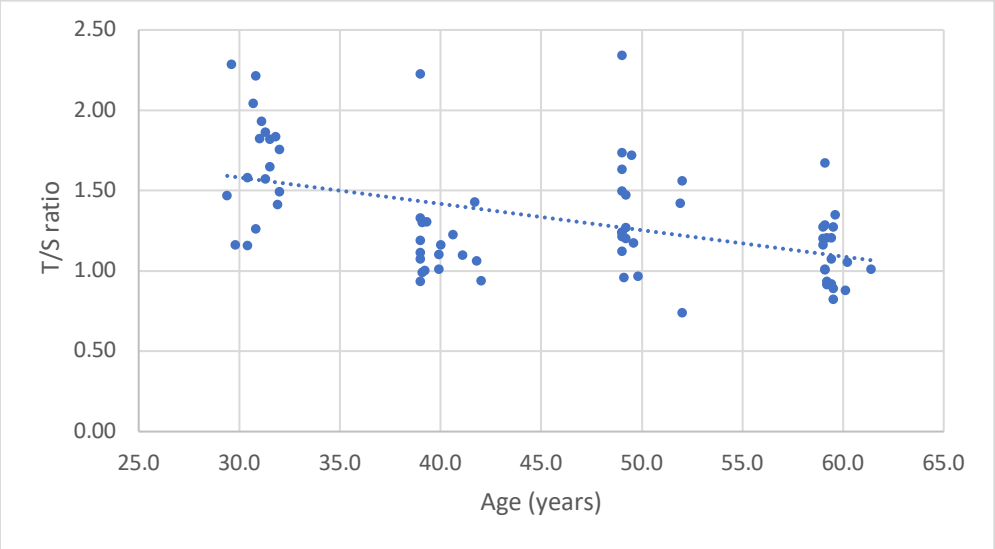


B. Fourteen samples measured at two phases with 3-5-year intervals.





C. Seventy-four samples measured at phase 1.



4.3.3 Longitudinal changing rates of LTL within long (8-10-year) and short (3-5-year) time intervals

The 40 samples with 8-10-year intervals showed a wide range of differences between two time-point measures (mean[SD]=-0.31[0.18]), ranging from -0.66 to 0.24, which corresponded to a LTL change of 39.7% decrease to 44.4% increase compared to the baseline LTL measures. All except one sample had shorter LTLs in phase 2 than in phase 1. Excluding the outlier that showed LTL lengthening, the range upper bound changed to -0.03, corresponding to a decrease of 3.8% of the baseline measure. Within this set of 40 samples, the shortening rates varied between individuals (mean[SD]=-0.04[0.02] per year, Figure 4.3), and were associated with baseline LTL measures (beta coefficient[SD] = -0.07[0.01], *p*-value=$6.47X10^{-8}$, Figure 4.5A), but not age (beta coefficient[SD] = 0.0002[0.0005], *p*-value=0.70, Figure 4.5B). The association between shortening rate and baseline LTL was robust after adjusting for age (beta coefficient[SD] = -0.08[0.01], *p*-value=$4.55X10^{-9}$). Excluding the one outlier that showed LTL lengthening, mean and SD of LTL shortening rates remained the same, as well as associations with baseline LTL with or without adjustment for age.

Similarly, twelve out of fourteen samples with 3-5-year intervals exhibited LTL shortening overtime. Differences between two time-point measures (mean[SD]=-0.12[0.12]) ranged from -0.29 to 0.14 overall, and to -0.02 after excluding the two outliers with LTL lengthening.

These corresponded to a LTL change from -25.0% to 14.0% of the baseline measures, and to -2.2% after excluding the two outliers. The shortening rate (mean[SD]=-0.03[0.03] per year) was not associated with age (Figure 4.5B), while the association with baseline LTL was marginally significant (beta coefficient[SD]=-0.09[0.04], *p*-value=0.06, Figure 4.5A), and remained at the same level of significance after adjusting for age. After excluding the two outliers of LTL lengthening, the association strength between shortening rate and age slightly increased (beta coefficient[SD]=-0.08[0.03], *p*-value=0.02), and stayed robust after adjustment for age.

We compared LTL shortening rates estimated using the 40 samples with longer (8-10-year) time intervals against those using 14 samples with shorter (3-5 years) time intervals. The average shortening rates within the two sets of time intervals were statistically comparable (Mann-Whitney U test *p*-value = 0.60, Figure 4.6), although longitudinal changes of LTL were smaller within shorter than longer time intervals (Mann-Whitney U test *p*-value = $7.12X10^{-5}$, Figure 4.6) as expected. These results were similar after excluding outliers that showed LTL lengthening.

Figure 4.5 Longitudinal changing rates of LTL within 3-5- and 8-10-year intervals.
Changing rates are stratified by quartiles of baseline LTL in the panel **A**, and by baseline age groups (30-40 years, 40-50 years and 50 years and above) in the panel **B**. The pale blue and red bars indicate 3-5-year (N=14) and 8-10-year (N=40) intervals, respectively. Within each boxplot, the horizontal line reflects the median, the top and the bottom of each box reflect the interquartile range, the whisker indicates an additional 1.5 times interquartile range within each grouping, and the red dots show outliers.

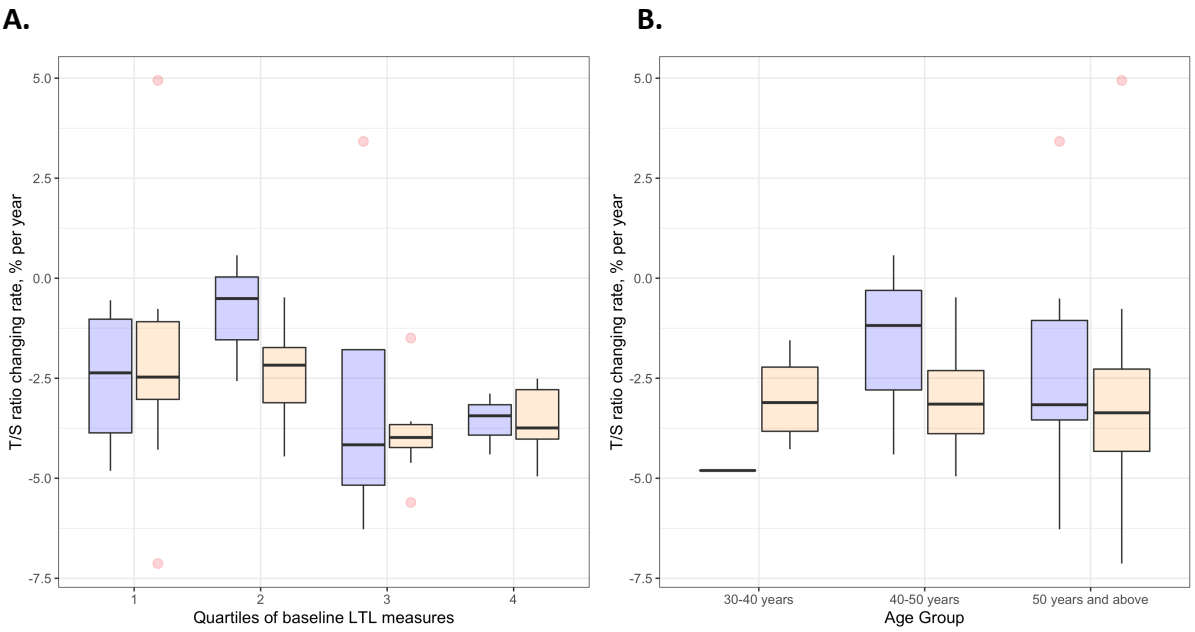
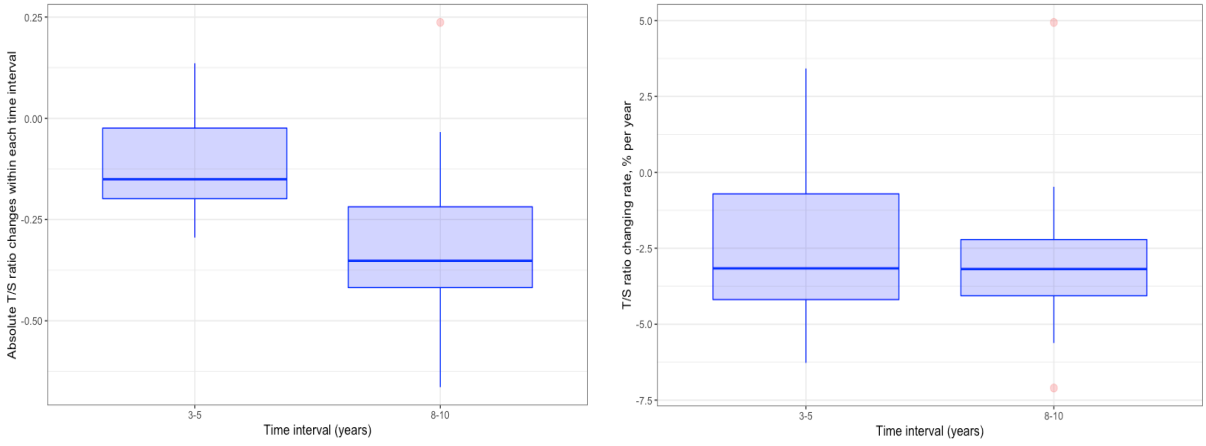


Figure 4.6 Comparing longitudinal changes and annual changing rates of LTL within 3-5- and 8-10-year intervals.

Boxplots are drawn in the same format as described above. The left and right panels show longitudinal changes and annual changing rates of LTL, respectively. The T/S ratio (LTL measure) change within each time interval is shown as the y-axis in the left panel and the T/S ratio (LTL measure) changing rate in % per year is shown as the y-axis in the right panel.



4.4 Summary and discussion

The work presented in this chapter aims to examine feasibility of conducting large-scale observational and GWAS on longitudinal changes of LTL. To examine such feasibility, I

conducted a systematic literature review, and performed pilot study analyses with two time-point measures of LTL in the Fenland prospective cohort.

4.4.1 Systematic literature review

LTL attrition rates have been suggested to increase with age and their association patterns with age vary between men and women. The baseline LTL measure has been consistently reported to be strongly associated with LTL attrition rates. Other environmental and behavioural factors may also influence the LTL attrition rates, but previous reports are not consistent. These factors include smoking, alcohol intake, dietary patterns, physical activity, body weight and other metabolic traits. To address such discrepancies, further studies are needed with larger sample sizes and statistical models that can properly address technical issues of repeated measures in longitudinal studies, such as the issue of regression to the mean. Furthermore, accelerated LTL attrition has been suggested to be associated with higher risks of age-related common complex diseases, such as cardiometabolic morbidities and mortality and haematologic malignancies. These associations with clinical outcomes show the promise of using LTL attrition rate as a biomarker to predict disease risk and progression, and patients' responses to therapeutic treatments. However, current studies, including both population cohort studies and clinical trials are far from being sufficient to draw any solid conclusions about clinical values of the LTL attrition rates. Further studies are needed to validate these associations and explore aetiological roles of LTL attrition in disease occurrence and development.

4.4.2 Pilot analyses within the Fenland study

In the pilot study, I found the longitudinal changing rate of LTL was associated with baseline LTL measurement as previously reported, but not age. The changing rates estimated within longer and shorter time intervals showed comparable results, indicating that duration of time intervals does not affect such estimation.

There are few cohort studies that have analysed longitudinal changes of LTL, with extensive heterogeneities in their association findings. In fact, most of the reported associations are inconsistent, except with baseline LTL. Therefore, a prospective cohort study with a larger sample size and optimised measuring approach can potentially help to

distinguish true association findings from false positive ones. However, there are plenty of technical and statistical issues in analysing longitudinal measures of LTL. First of all, regression to the mean, a statistical phenomenon that is commonly observed in repeated data, where relatively high (or low) measured values are likely to be followed by less extreme ones near the true means of these values in the same individuals[352]. This issue is caused by non-systematic variation (i.e. random errors) in the measurements of LTL, and practically can cause a problem in distinguishing a true change from an expected change due to natural fluctuation of the data. Because of this issue, participants with much longer LTL measurements in the phase 1 will have greater decreases (i.e. faster shortening) of LTL measurements in the phase 2, whereas those with shorter LTL measured in the phase 1 tend to have less decreases or even elongation (i.e. slower shortening or lengthening) of LTL in the phase 2. In support of this possibility, we and others have consistently found LTL at baseline is strongly positively associated with LTL shortening rate. Therefore, the regression to the mean problem must be taken special care of in the future longitudinal studies of LTL. Secondly, it can be technically challenging to discriminate between LTL measures at baseline and longitudinal changes of LTL, as these two are highly correlated. Adjusting for baseline LTL in regression models that test associations with LTL shortening rates may induce an issue of collinearity, and thus challenges model stability. Thirdly, a selection bias of samples may be present when measuring LTL in the phase 2. Because as previously reported, various clinical disorders are associated with shorter LTL (section 1.4.1.4), these disorders may reduce chances of participants being re-examined in the second visit, causing individuals with relatively shorter LTL to become largely underrepresented in the repeated measures.

GWAS can help to identify genetic variants that are strongly associated with LTL attrition rates, which can be useful in delineating causality in the MR framework. Given that epidemiological studies of LTL shortening so far have all been observational, therefore, associations identified in these studies can be confounded by known and unknown risk factors and biased by reverse causality. Causal associations between previously identified risk factors and diseases and LTL shortening rate are still unexplored. MR analyses leveraging genetic instruments as proxies for the LTL shortening rate may assist in dissecting causal pathways. Therefore, a primary stimulus of performing the GWAS and thus this feasibility study is to identify genetic factors that are strongly and specifically associated with LTL shortening, which could then be used as instruments for MR-based causal association analyses. However,

identifying robust genetic instruments for LTL shortening can have several challenges: 1) There may be insufficient power to discover any variants due to relatively small sample sizes. Additionally, external studies with both longitudinal LTL changes and genotypes measured, such as the PREVEND study, have the capability of performing GWAS in parallel to our study, and can potentially be meta-analysed with our study to increase overall discovery power. (2) Even if there are variants found to be strongly associated with LTL shortening rates, it is technically challenging to determine whether these variants are specifically associated with LTL shortening or simply baseline LTL, as these two are highly correlated, both observationally and biologically. Therefore, even larger sample sizes with more advanced mathematical modelling are required to conduct such GWAS.

Even if the GWAS on LTL shortening rate lacks power to identify any genome-wide significant loci, the summary statistics produced from these studies can still be useful for reverse MR analyses, in which causal effects of diseases or intermediate traits on LTL shortening can be tested. For example, genetic instruments for cardiometabolic traits and diseases have been well established and the their summary statistics are publicly available, by linking their association estimates with those with LTL shortening rates in the two-sample MR framework, we can shed light upon whether and how cardiometabolic diseases or the related risk factors influence LTL shortening.

In conclusion, studies on longitudinal changes of LTL can be scientifically interesting, but technically challenging. To tackle the challenges, larger sample sizes with more advanced statistical modelling may help.

Chapter 5

Characterisation of mLOY and its association with T2D risk

Abstract

Background mLOY, the most common post-zygotic chromosomal alteration in men, has been strongly correlated to age and age-related common complex diseases. While mLOY has been suggested to play an essential role in cancer development, few studies have analysed prospective associations of mLOY with T2D risk.

Objective To characterise observational associations between mLOY and T2D.

Methods mLOY was estimated based on SNP-array intensity data and expressed as the median value of logarithmic ratios of observed to expected intensity values (R, $\log_2(R_{observed}/R_{expected})$, mLRR) across all SNPs on chromosome Y specific regions. mLRRY was analysed as a continuous variable with mLRRY<0 indicating mLOY. Association between mLOY and incident T2D risk was analysed in the EPIC-InterAct case-cohort study (n=11,892 men, 51.84% cases), using Prentice-weighted Cox regression models with age as the underlying timescale and adjusted for age, sex, smoking and other lifestyle factors. The association was analysed in each country separately, with results meta-analysed using random-effects models. UK biobank (221,597, 3.38%) was used as a replication cohort.

Results Men with increased mLRRY (less mLOY) were at a modestly lower risk of T2D after adjusting for age, centre and genotyping array, but the association was not statistically significant (HR[95%CI] = 0.91[0.83-1.01] per 1-SD increased mLRRY, *p*-value=0.07), and showed large heterogeneity across countries (I^2=81.7%). Adjusting for smoking further attenuated the association (HR[95%CI] = 0.94[0.86-1.04], *p*-value=0.23). Younger men (<50 years) exhibited larger risk effects (HR[95% CI] = 0.95[0.91-0.99], *p*-value=0.02) compared to older men (50-65 years: HR[95%CI] = 0.98[0.94-1.02], *p*-value=0.33; and >65 years: HR[95%CI] =0.91[0.79-1.04], *p*-value=0.18; $p_{interaction}$=0.002). Other than smoking, no modifiable risk factors showed significant associations with mLRRY after Bonferroni correction (*p*-value=$1.52X10^{-3}$, for 33 traits tested). In UK biobank, the association between mLRRY and T2D risk was significant but in the opposite direction (HR[95%CI] = 1.05[1.03-1.07], *p*-value=$4.27X10^{-5}$). The association significance was reduced in men younger than 50 years (HR[95%CI] = 1.08[0.99-1.16], *p*-value=$8.84X10^{-2}$), and there was no evidence for interaction between mLRRY and age for their effects on the T2D risk ($p_{interaction}$=0.43).

Conclusion Observational evidence shows no strong support for an association between mLOY and T2D risk, considering the roles of confounding and inconsistency of results across study cohorts.

5.1 Introduction

mLOY in peripheral blood is the most common mutation acquired during adulthood for men[353]. The occurrence of mLOY is strongly correlated to age, with ~20% of men aged 80 years or older having >10% of blood cells with mLOY[193,207].

mLOY in leukocytes has been suggested as a signal of impaired immunosurveillance, leading to disrupted immune response, thereby increasing risks of tumorigenesis in various tissues, neurodegenerative development, CVD, T2D and autoimmune disorders[354–356]. Strong relationships have been established between mLOY and cancer diagnosis and mortality[193,196,357–359], but few studies have been undertaken to investigate the role of mLOY in cardiometabolic disorders[356,357,360].

T2D has been identified as a major risk factor for mortality and a wide range of clinical disorders that influence human longevity, including CVD and cancers[361]. It has grown rapidly to epidemic proportions worldwide and deemed as an essential cause for accelerated ageing. It has been suggested that glucose homeostasis plays an important role in regulating life span in animal models; in humans, genetic variations in genes involved in GH and IGF1 signalling pathways are associated with longevity (section 1.3.2)[114,362,363]. Besides conventional biomarkers for T2D risk prediction, including family history, obesity, blood pressure, and HbA1c[364,365], recent developments in omic biomarkers have demonstrated some potential values in predicting T2D risk, including PRS[366,367] and branched chain amino acids[368–372]. However, the role of mLOY in T2D susceptibility has not been fully understood. Here, using the largest incident T2D case-cohort study, the EPIC-InterAct, we were able to examine associations of mLOY with T2D incidence and shed light upon potential utilities of mLOY in early and improved risk assessment of T2D.

5.2 Methods

5.2.1 Population

5.2.1.1 EPIC-InterAct

InterAct is a case-cohort study nested within the EPIC cohort, designed to investigate how inherited and modifiable risk factors interact to influence T2D susceptibility. Participants and study design have previously been described[212]. In brief, the project involves 23 research centres across Europe in 8 countries (Denmark, France, Germany, Italy, Netherlands, Spain, Sweden and UK). All incident T2D occurring in the EPIC cohort between 1991 and 2007 were ascertained using multiple sources of evidence (self-report, linkage to primary care registers, secondary care registers, medication use, hospital admission data, and mortality data). In total, 340,234 EPIC participants were included in the InterAct, and followed up for a mean [range] of 11.7 [0–17.5] years, during which 12,403 incident T2D cases were verified[213]. A centre-stratified, random sub-cohort of 16,835 individuals was selected, among which 778 individuals developed incident T2D. Only men were included (n=12,238, from 7 countries), and after excluding those who had missing mLOY measurements, there were 6,099 men left. Missingness of mLOY was further examined with regards to missing patterns and random distributions (Supplementary Table 22). We further removed men who showed any evidence of diabetes at baseline, had no blood samples stored, or diabetes status missing, which resulted in a total of 5,841 men included for analyses. All participants gave written informed consent, and the study was approved by local ethics committees in the participating countries and the Internal Review Board of the International Agency for Research on Cancer.

5.2.1.2 UK Biobank

We used data collected as part of UK biobank, a prospective cohort study of over 5 million individuals aged 40-69 years and recruited from 22 assessment centres across England, Scotland and Wales[302]. Participants provided baseline information on demographic, lifestyle and other health-related factors through online questionnaires and completed a range of physical and imaging-based measurements. They also provided biological samples which allowed various biochemical, genomic and other omics-based assays to be conducted, and objective measures of physical activity and multi-modal imaging assessment within different subsets of the cohort. The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee and all participants provided written informed consent. To define T2D cases, we implemented a previously calibrated algorithm to call prevalent T2D cases, which comprised self-reported and nurse interview validated diagnoses, ages at diagnoses

and diabetes medications and complications (n=487,915, 4.08% cases)[373]; and incident T2D cases (2.42%) using heath-care data linked from hospital episode statistics and cause of death data from the National Death Registries. Disease categories were defined using the ICD10 codes (E11: T2D). Participants whose dates of their first hospital admissions for T2D preceded the baseline assessment dates or with HbA1c ≧ 48mmol/mol at baseline were excluded. Samples from the full release of UK biobank (May 2018) were analysed for estimating mLOY. In addition to the centrally performed QC procedures by UK biobank[303], we further excluded individuals who shared relatedness closer than third degree and of ancestry other than white European. We restricted our analyses to men (n=221,597).

5.2.2 Genotyping and mLOY measurements

Blood samples of 10,004 men in EPIC-InterAct were genotyped on two slightly different arrays: About 61% (6,102) were analysed on the Infinium CoreExome-24 v1.3 BeadChip array and the remainder (39%, 3,902) on the HumanCoreExome-12 v1.1 BeadChip array. Imputation was performed according to the HRC panel[14]. Genotyping, imputation and mLOY measurements in UK Biobank were previously described elsewhere[195,357,374].

5.2.2.1 Continuous and binary measurements of mLOY

mLOY of each individual was measured using SNP microarrays-based method coupled with the PennCNV calling algorithm[375]. Observed signal intensity values (R) were estimated using SNP array-intensity data and shown as average read depths (normalized signal intensities) over chromosome Y, exclusively the X-degenerate regions[193,195,196,207]. Expected R values were extracted from a standard reference (*.egt) file. The ratios of observed to expected R values on a base 2 logarithmic scale ($\underline{\log_2}(\underline{R}_{observed}/\underline{R}_{expected})$, LRR) were calculated for all SNPs that passed genotyping QC and showed bi-allelic R patterns within the male-specific region of chromosome Y[375].

For each participant, the median value of LRRs of all SNPs within the male-specific region of chromosome Y (mLRRY), was used as the quantitative measurement of mLOY, with negative values indicating LOY. In analyses using binary measurements of mLOY, the variable was defined as “mLOY” based on mLRRY<0.

5.2.2.2 Distributions of mLRRY and data transformation

mLRRY values ranged from -4.6 to 0.3 in the EPIC-InterAct sub-cohort and -0.3 to 0.3 in the UK biobank. Distributions of the mLRRY values were examined overall, and in each country of the EPIC-InterAct study, separately. They exhibited normal distributions with means centred at 0 and SDs ranging from 0.04 to 0.14 (Supplementary Table 18, Supplementary Figure 4 and 5). However, normality tests of mLOY measures showed large absolute values of skewness (-23.1) and kurtosis (935.3) in the EPIC-InterAct, indicating a heavy-tailed, asymmetric distribution, i.e. left-skewed deviation from normality. Therefore, to reduce effects of potential outliers that constituted the heavy tails, and maintain consistency between analyses within the two cohorts, the following data processing procedures were performed in both cohorts, (1) winsorisation of mLRRY values at 5SD to exclude extreme outliers; (2) inverse normal transformation separately for each genotyping array; and (3) standardisation to a distribution with a mean of 0 and a SD of 1.

5.2.3 Covariates

Besides age (age at recruitment), study centre and genotyping array, other covariates were considered for inclusion based on previous evidence for their associations with T2D, which were illustrated as below. *Smoking*: smoking status was classified into four categories: never, former and current smokers and unknowns; or two categories: never and ever smokers. *Alcohol*: categorised into six groups (never: 0 g/day, light: 0.1-4.9 g/day, moderate: 5-14.9, regular: 15–29.9, heavy: 30–59.9, and extreme ≥60 g/day drinkers)[376]. *Education*: educational levels were self-reported the highest and categorized into five categories (with the unspecified excluded): none, primary school, technical school, secondary school, university degree[377]. *BMI*: BMI was measured (kg/m^2) with correction for clothing and assessed as a continuous variable. *Waist circumference*: Waist circumference was measured either at the narrowest circumference of the torso or at the midpoint between the lower ribs and the iliac crest, assessed as a continuous variable[378]. In the EPIC-InterAct, two additional covariates were included in the fourth model, which were described as follows. *Mediterranean diet score*: dietary intake of nine nutritional components characteristic of the Mediterranean dietary

pattern was estimated (gram per 1,000 kcal, except alcohol consumption) and each divided into 3-quantiles, according to the distributions observed in the EPIC-InterAct subcohort[379]. A score was derived by adding up these 3-quantiles, therefore ranging from 0 to 18, which were further classified into three categories (low: 0-6 points, medium: 7-10 points, high: 11-18 points). *Physical Activity*: physical activity levels were assessed at baseline by a validated self-report questionnaire combining occupational and leisure time physical activity levels, and categorized into four groups (inactive, moderately inactive, moderately active and active) according to the Cambridge Physical Activity Index[380].

5.2.4 Statistical analyses

To estimate associations between mLRRY (mLOY measurements) and T2D risk in the EPIC-InterAct cohort, I applied Prentice-weighted Cox regression models modified for case-cohort analyses with age as the underlying timescale within each country[212,292], and HRs were combined using random-effects meta-analyses. The percentage of overall variation in the HRs attributed to heterogeneity between countries was estimated and expressed as I^2. In the UK Biobank cohort, we used multivariable logistic regression models to examine risk effects of mLRRY on prevalent T2D, and Cox regression models for incident T2D. In either cohort, four models with different adjustments were considered. The first model included adjustments for genotyping array and centre, the second for one additional adjustment for age, the third for two additional adjustments for age and smoking status, and the fourth for multiple factors including age, smoking status, lifetime alcohol consumption, educational level, BMI and waist circumference within both cohorts, and Mediterranean diet score and physical activity level in the EPIC-InterAct, as only a small subset of individuals in the UK Biobank had equivalent measures for these two covariates. The main exposure, mLRRY, was normalised and standardised as described in the 5.2.2.2, and treated as a continuous variable in the primary analyses, and in the secondary analyses, dichotomised by 0 (men with mLRRY<0 as mLOY cases, coded as 0; whereas men with mLRRY≧0 as mLOY controls, coded as 1).

To assess linearity of associations between mLRRY and incident T2D risk in the EPIC-InterAct, I further analysed the primary exposure of mLRRY in quantiles. These analyses were conducted in each country separately using Prentice-weighted Cox regression models as described above; the resultant HRs were combined using the inverse-variance weighted random-effects meta-analysis models within each quantile.

To assess interactions of mLRRY with age or smoking on their risk effects on T2D, I conducted stratification analyses. Age was stratified into three groups (<50, 50-65 and ≧65 years old) in the EPIC-InterAct and five groups (<50, 50-59, 60-69, ≧70) in the UK biobank, and the associations of mLRRY with T2D risk were analysed within each age group, and in the EPIC-InterAct, additionally in each country as well. Similarly, smoking status was stratified into two (never or ever smokers) groups, with subjects with missing smoking status excluded, and associations of mLRRY with T2D risk were analysed within each smoking status group in each country of EPIC-InterAct or in UK biobank. In the EPIC-InterAct, HRs estimated across countries were meta-analysed using random-effects models, and the resultant country-combined HRs were further meta-analysed using fixed-effects models across age or smoking strata. Heterogeneities across age or smoking strata were evaluated using I^2 estimates. In addition, significance levels of interactions between mLRRY and age or smoking groups were assessed using likelihood ratio tests.

To identify additional risk factors for mLOY, I performed an exploratory analysis, examining associations between mLRRY and a variety of traits, including lifestyle and anthropometry traits and circulatory biomarkers in the quasi-random subcohort of the EPIC-InterAct study. Linear regression models were applied with adjustments for age, centre and genotyping array, in each country separately, and the resulting beta estimates were meta-analysed across countries using random-effects models. Continuous factors were normalised by inverse normal transformation and standardised to distributions with means of 0 and SD of 1. Categorical factors were analysed at an ordinal scale and treated as continuous variables to avoid sparsity in certain strata.

5.3 Results

5.3.1 Baseline characteristics of mLOY measurements

Characteristics of the two study population cohorts, EPIC-InterAct case-cohort and UK biobank were presented overall and stratified by mLOY indicator (mLRRY<0, ≧0, or missing, Table 5.1). In EPIC-InterAct, a total number of 6,099 men with non-missing mLRRY estimates were included in the analyses, among which 1,413 men had detectable mLOY (mLRRY<0) and of these men 73 had relatively higher degrees of mLOY (mLRRY<-0.15). In UK biobank, all men (n=221,597) of white European ancestry had mLRRY measured, among which 44.7% had detectable mLOY (mLRRY<0) and 8.24% showed higher degrees of mLOY (mLRRY<-0.15). The proportions of men with college/university education were comparable between the three groups of mLOY measurements (positive, negative and missing mLRRY values) in either the EPIC-InterAct T2D case or the quasi-random sub-cohort, or in the UK biobank. Adiposity levels were also similar between the three groups of mLOY measurements in the three study parts (Table 5.1).

Previous studies have identified age and smoking as strong risk factors for mLOY[193,196,207,353], and these two factors have also been reported to affect T2D risk[212,213,381], therefore we considered them as potential confounders, and investigated their associations with mLOY (mLRRY<0). In line with previous studies, I found the prevalence of mLOY (mLRRY<0) was relatively higher in men at older ages (≧65 years) or among ever (current and previous) smokers (Table 5.2). Higher degrees of mLOY (mLRRY<-0.15) were only present among elderly men above certain ages (50-60 years in EPIC-InterAct and 40-50 years in UK biobank), with the prevalence increasing along with age, reaching 7.87% and 5.48% among men over 70-year old in EPIC-InterAct and UK biobank cohorts, respectively (Table 5.2). In contrast to age, the higher degrees of mLOY (mLRRY<-0.15) were present in all smoking status, but more frequently observed in current or previous smokers than never smokers (Table 5.2).

Table 5.1. Baseline characteristics of the study population cohorts, overall and stratified by the mLOY indicator.

mLOY indicator: mLRRY≧0, <0 or missing (.). Values represent means (SDs) for continuous variables and percentages for categorical variables.

	EPIC-InterAct									UK Biobank		
	Overall	T2D case				subcohort				total	mLRRY≧0	mLRRY<0
		total	mLRRY≧0	mLRRY<0	mLRRY=.	total	mLRRY≧0	mLRRY<0	mLRRY=.			
No.participants	11,892	5,781	1,965	683	3,133	6,111	2,488	704	2,919	221,597	122,597	98,984
Incident T2D, %	51.84	100.00	100.00	100.00	100.00	6.28	5.39	3.27	7.78	3.38	3.08	3.77
College/university education, %	20.27	16.52	17.46	16.54	15.93	23.81	25.32	25.28	22.17	33.78	34.61	32.77
Age, years	54.1 (8.4)	55.4 (7.5)	55.1 (6.8)	58.0 (7.3)	55.0 (7.9)	52.9 (8.9)	52.4 (8.6)	56.1 (8.6)	52.4 (9.1)	56.8 (8.2)	55.6 (8.2)	58.1 (7.9)
Body mass index, kg/m^2	27.9 (4.1)	29.3 (4.1)	29.4 (4.2)	28.7 (3.9)	29.4 (4.0)	26.6 (3.6)	26.5 (3.6)	26.2 (3.4)	26.8 (3.6)	27.8 (4.2)	27.9 (4.3)	27.7 (4.2)
Waist circumference, cm	98.8 (11.0)	102.6 (10.6)	103.0 (10.7)	102.0 (10.6)	102.5 (10.4)	95.1 (10.1)	95.0 (10.2)	94.6 (9.7)	95.3 (10.1)	96.3 (11.3)	96.9 (11.4)	96.9 (11.2)
Age group												
<50 years, %	28.81	22.44	17.91	12.59	27.42	34.84	31.99	19.74	40.90	21.75	27.59	14.52
50-65 years, %	64.31	70.11	77.96	75.11	64.09	58.83	64.47	68.75	51.63	51.34	55.88	45.73
≧65 years, %	6.88	7.46	4.12	12.30	8.49	6.33	3.54	11.51	7.47	18.30	16.54	20.49
Smoking status												
Current smoker, %	32.43	33.71	33.94	42.90	31.57	31.22	29.90	36.36	31.11	12.33	11.46	13.42
Previous smoker, %	38.40	40.72	41.17	38.80	40.86	36.21	36.94	37.36	35.32	38.26	37.12	39.68
Never smoker, %	27.85	24.23	24.07	16.98	25.92	31.27	31.83	24.86	32.34	48.87	50.89	46.38
Unknown, %	1.31	1.33	0.81	1.32	1.66	1.29	1.33	1.42	1.23	0.53	0.54	0.52

Table 5.2. mLOY distribution, overall and stratified by 10-year age bin and smoking status.
Counts and frequencies of mLOY (mLRRY<−0.15 or <0) were calculated in each stratum.

	EPIC-InterAct					UK Biobank				
	Total count	mLRRY < 0		mLRRY < -0.15		Total count	mLRRY < 0		mLRRY < -0.15	
		Count	Frequency	Count	Frequency		Count	Frequency	Count	Frequency
Age band										
20-30	53	7	13.21%	0	0%	0	0	0%	0	0%
30-40	195	28	14.36%	0	0%	6	2	33.33%	0	0%
40-50	1,184	190	16.05%	0	0%	51,623	17,802	34.48%	8	0.02%
50-60	3,035	603	19.87%	18	0.59%	70,960	29,506	41.58%	299	0.42%
60-70	1,501	505	33.64%	45	3.00%	97,841	51,003	52.13%	2271	2.32%
70-80	127	78	61.42%	10	7.87%	1,167	671	57.5%	64	5.48%
Smoking status										
Never smoker	1,624	296	18.23%	9	0.55%	108,295	45,908	42.39%	743	0.69%
Previous smoker	2,361	534	22.62%	29	1.23%	84,792	39,281	46.33%	1264	1.49%
Current smoker	2,039	561	27.51%	34	1.67%	27,326	13,279	48.59%	616	2.25%
Unknown	71	20	28.17%	1	1.41%	1,184	516	43.58%	15	1.27%

5.3.2 Observational associations of mLOY measures with T2D risk

5.3.2.1 EPIC-InterAct

Men with decreased mLRRY values were at a suggestively higher risk of T2D (HR[95%CI] = 0.91[0.83-1.00], *p*-value=0.05, Figure 5.1, Supplementary Figure 6) in the basic model adjusted for centre and array. While additionally adjusting for age (main model) had little impact on the association estimate (HR[95%CI] = 0.91[0.83-1.01], *p*-value=0.07, Figure 5.1, Supplementary Figure 6), adjusting for both age and smoking (model 3) markedly attenuated the association strength (HR[95%CI] = 0.94[0.86-1.04], *p*-value=0.23, Figure 5.1, Supplementary Figure 6). Further adjustment for multiple additional covariates (model 4) did not change the association estimate as much as for age and smoking in the model 3 (Figure 5.1, Supplementary Figure 6). All these models exhibited large values of I^2, indicating substantial levels of heterogeneity between countries. Moreover, given that the basic model produced a relatively stronger result than the other models, to eliminate the possibility that it is merely because of the slightly larger sample size in this model than in the other models, I restricted the analysis to the complete set of 5,191 men within whom the model 4 was performed. As a result, the association was comparable (HR[95%CI] = 0.91[0.83-1.00], *p*-value=0.05, 51.6% cases), suggesting that the attenuated association results in the other models compared to the basic model were mainly because of the adjustment for confounding factors rather than the sample size. Applying the same models but using a binary variable of the mLOY indicator (mLRRY<0) as the main exposure showed similar results, but they were all non-significant (Supplementary Figure 7).

5.3.2.2 UK Biobank

In line with the previous study, we found a negative association of mLRRY with the risk of prevalent T2D[373] (OR[95%CI] = 0.79[0.78-0.81], *p*-value=0.013) in a model consistent with the previous paper[357], which only adjusted for two covariates (centre and genotyping array in UK biobank). Interestingly, additional adjustment for age substantially increased the strength of the association, but led to an reversal of the direction of the association (OR[95%CI] = 1.08[1.06-1.10], *p*-value=$2.33X10^{-16}$). Given that age is a confounding factor, showing strong associations with both the main exposure (mLRRY) and the outcome (T2D), such models should include age in their adjustments.

Next I investigated why adding age as a covariate completely changed the direction of the association. I suspected that this might be biased by a multicollinearity problem. In the model that adjusted for age, centre and genotyping array, age demonstrated a large variance inflation factor (48.86), indicating a major concern of collinearity, even though the pair-wise correlation between age and mLRRY was small (Spearman's Rho = -0.2). Further, the condition number was extremely large (978.2), indicating a global instability of beta coefficients estimated from these models in UK biobank.

Association with incident T2D was more significant than that found in the EPIC-InterAct, but in an opposite direction, either with or without adjustment for age (Table 5.3, Supplementary Table 19). Additional adjustment for smoking alone or in combination with other confounding factors, including alcohol consumption, education level, BMI and waist circumference did not substantially change the result (Table 5.3, Supplementary Table 19). This discrepancy found between EPIC-InterAct and UK biobank studies is discussed in the section 5.4.4.

Figure 5.1. Observational associations between mLRRY and T2D risk across countries in EPIC-InterAct.

Associations were analysed using Cox regression models with different adjustments. The main exposure, mLRRY, was analysed as a continuous variable, with higher values indicating less 'loss' of Y. Association estimates across countries were combined using inverse variance weighted random-effects models, with between-country heterogeneity in each model quantified as I^2. Association estimates in the model adjusting for age, centre and array were shown in each country separately and combined, and in other models only summary association estimates were shown.

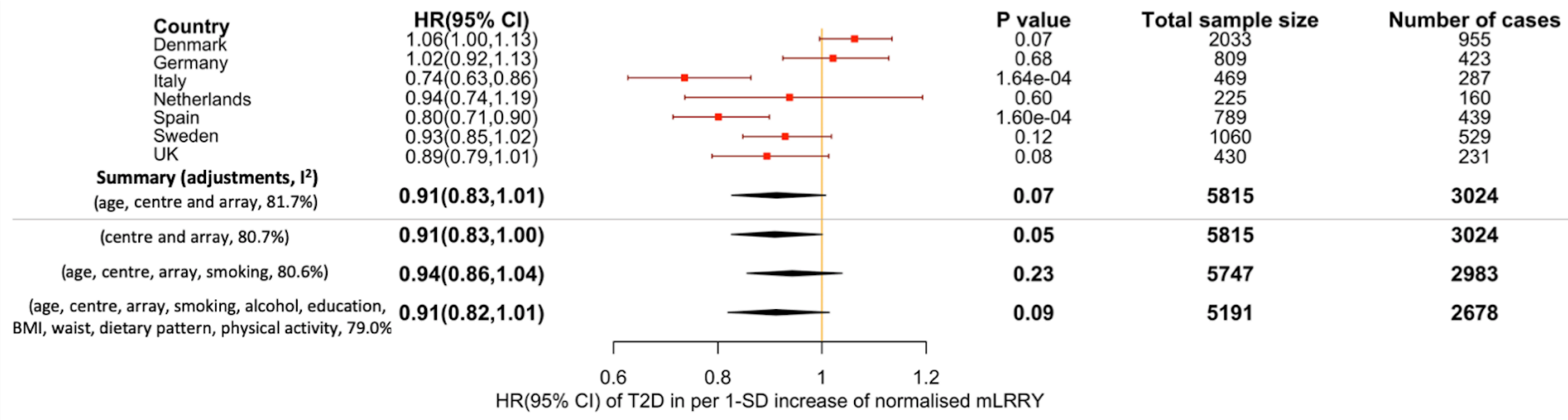


Table 5.3. Observational associations between mLRRY and T2D risk in UK Biobank.

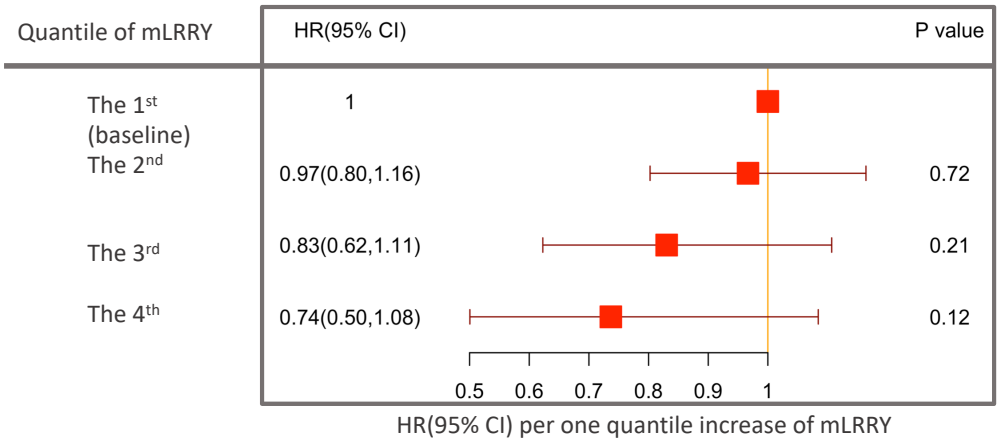
Associations were analysed using Cox or logistic regression models for incident and prevalent T2D, respectively, and with different adjustments, as shown in the table.

	Incident T2D (Cox regression models)						Prevalent T2D (logistic regression models)					
Adjustment	**HR**	**Beta**	**SE**	**P-value**	**total N**	**case N**	**OR**	**Beta**	**SE**	**P-value**	**total N**	**case N**
centre, array	1.06	0.06	0.01	5.22E-06	196,171	6,831	0.98	-0.02	0.01	1.28E-02	218,665	12,490
age, centre, array	1.05	0.05	0.01	4.27E-05	196,171	6,831	1.08	0.07	0.01	4.72E-16	218,665	12,490
age, smoking, centre, array	1.06	0.06	0.01	1.26E-06	195,992	6,822	1.08	0.08	0.01	1.76E-17	218,428	12,462
age, smoking, alchol consumption, education, BMI and wasit circumference, centre, array	1.04	0.04	0.01	5.88E-04	195,172	6,757	1.07	0.06	0.01	2.84E-11	217,239	12,357

5.3.3 Linear trend of associations between mLRRY and T2D risk

There was no evidence observed for a non-linear association between mLRRY and T2D risk. The linearity was examined by comparing effects of quartiles of mLRRY on T2D risk. The risk effects of mLRRY on T2D gradually increased in proportion to the increasing order of the quartiles of mLRRY, yet below nominal significance, which was possibly due to limited statistical power (Figure 5.2). The linear trend of associations between mLRRY quartiles and T2D risk was supported by the likelihood ratio test that showed low probabilities of being non-linear ($P_{non\text{-}linearity}$>0.5).

Figure 5.2. Association of quartiles of mLRRY with T2D risk.
The lowest quartile is set as the baseline. Quartiles were defined based on cut-offs derived in the InterAct subcohort (1st=[-3.42,-0.69], 2nd=[-0.69,-0.01],3rd=[-0.01-0.66], 4th=[0.66-3.42]).



5.3.4 Stratified analyses by age group or smoking status

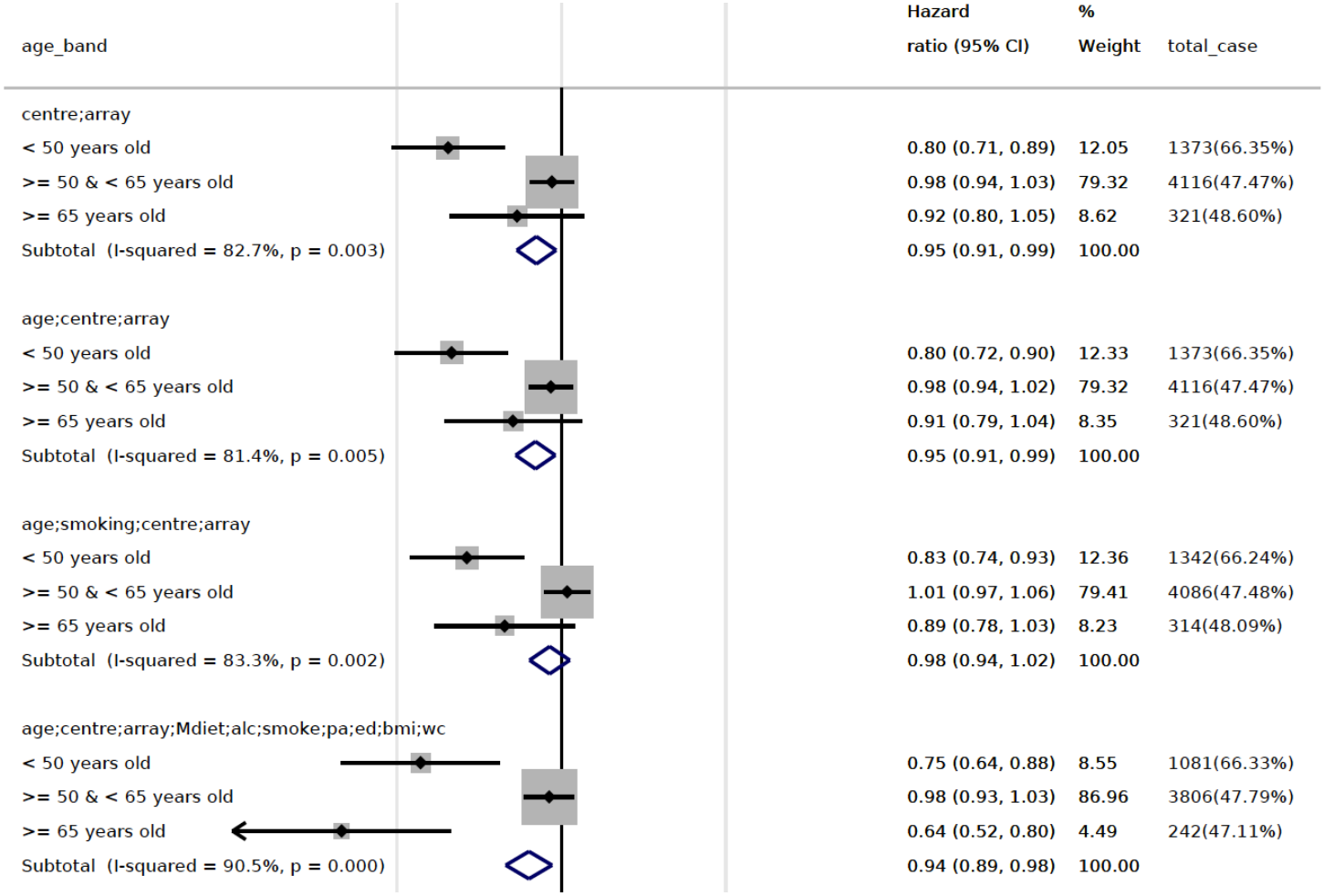
The risk effects of mLRRY on T2D differed between age groups, and the heterogeneity was consistently observed across all four models (I^2>80%, $P_{heterogeneity}$<0.005, Figure 5.4A). Individuals at younger ages (<50 years) exhibited higher risks of T2D per 1-SD decrease of mLRRY, which were attenuated to below nominal significance in men at older ages (≧50 years, Figure 5.3A). The variation in the risk effects of mLRRY on T2D across age bands was supported by a significant interaction observed between age-band and mLRRY ($P_{interaction}$<0.002). This suggested that mLOY might exert a relatively higher risk on T2D within the younger age group (< 50 years), even though the overall risk effect was weak.

Stratification by smoking status dampened the overall risk effect, with minor heterogeneity between different smoking strata (I^2<10%, $P_{heterogeneity}$>0.20, Figure 5.3B). Likelihood ratio tests also suggested a non-significant interaction between smoking status and mLRRY for their effects on the T2D risk ($P_{interaction}$>0.70).

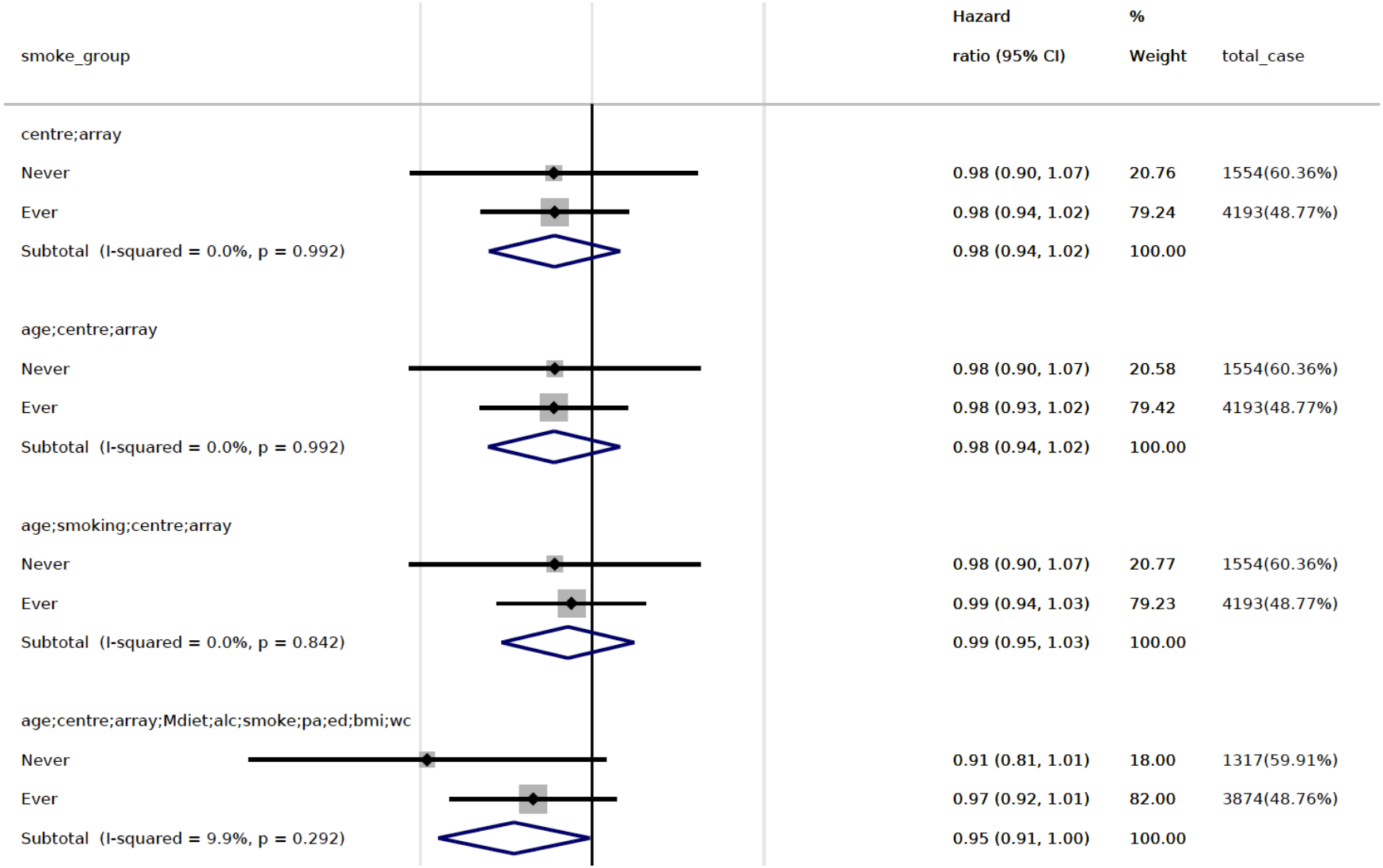
Figure 5.3. Stratified analyses by age group or smoking status.

A. Age and **B**. smoking status were divided into three groups with cut-offs at 50 and 65 years and two groups of never and ever smokers, respectively. Associations were analysed in each age group or smoking status in each country, and combined across countries using random-effects meta-analysis models. The country-combined association estimates in each age group or smoking status were further meta-analysed using fixed-effects models. The unit of X-axis is HR of T2D per 1-SD reduction of mLRRY, and the vertical line indicates HR=1.

A.



B.

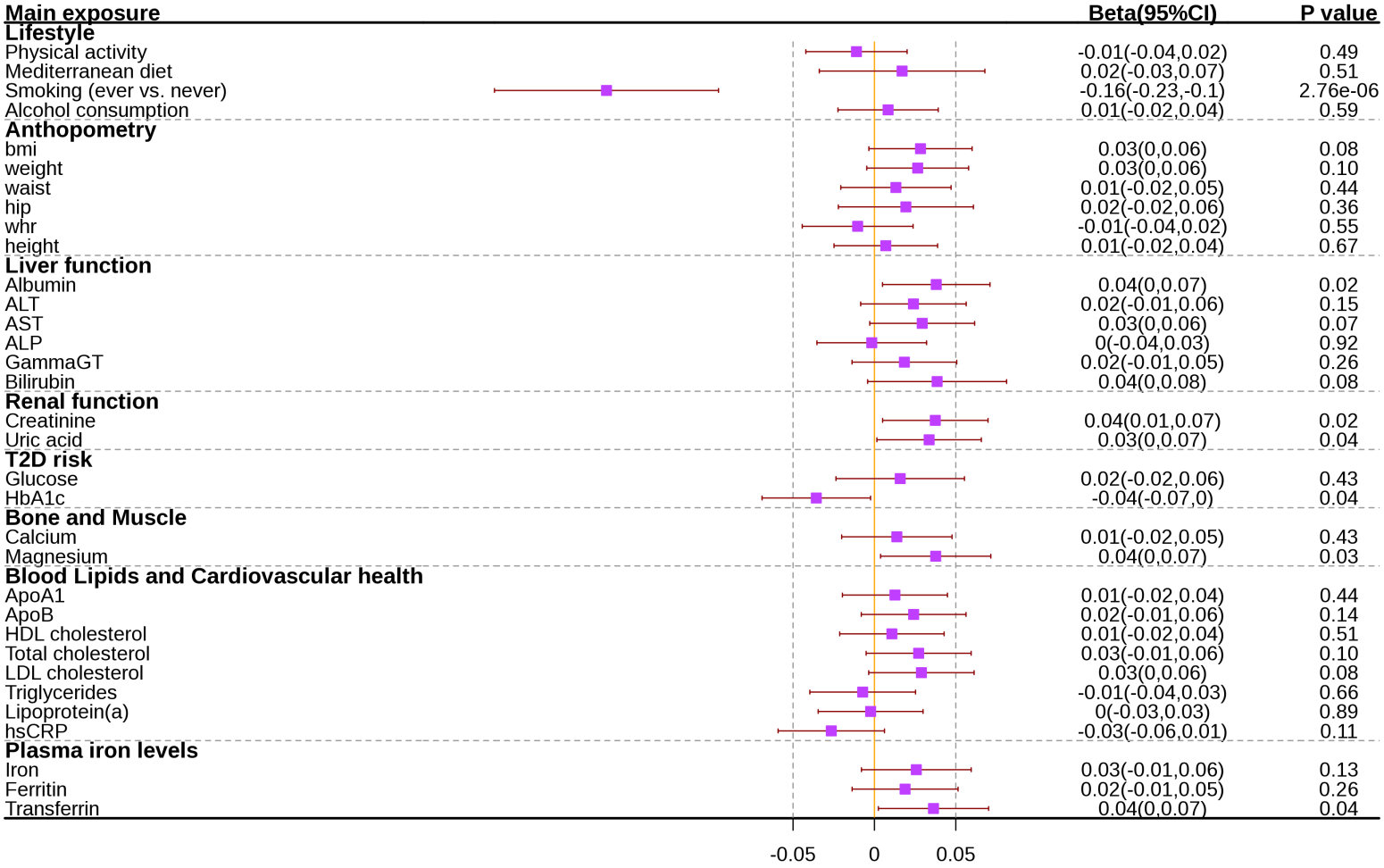

smoke_group
Hazard ratio (95% CI)
% Weight
total_case
centre;array
Never
0.98 (0.90, 1.07)
20.76
1554(60.36%)
Ever
0.98 (0.94, 1.02)
79.24
4193(48.77%)
Subtotal (I-squared = 0.0%, p = 0.992)
0.98 (0.94, 1.02)
100.00
age;centre;array
Never
0.98 (0.90, 1.07)
20.58
1554(60.36%)
Ever
0.98 (0.93, 1.02)
79.42
4193(48.77%)
Subtotal (I-squared = 0.0%, p = 0.992)
0.98 (0.94, 1.02)
100.00
age;smoking;centre;array
Never
0.98 (0.90, 1.07)
20.77
1554(60.36%)
Ever
0.99 (0.94, 1.03)
79.23
4193(48.77%)
Subtotal (I-squared = 0.0%, p = 0.842)
0.99 (0.95, 1.03)
100.00
age;centre;array;Mdiet;alc;smoke;pa;ed;bmi;wc
Never
0.91 (0.81, 1.01)
18.00
1317(59.91%)
Ever
0.97 (0.92, 1.01)
82.00
3874(48.76%)
Subtotal (I-squared = 9.9%, p = 0.292)
0.95 (0.91, 1.00)
100.00

5.3.5 Associations of mLOY measures (mLRRY) with lifestyle and anthropometry traits and circulatory biomarkers

Besides smoking, we investigated whether there were other modifiable risk factors associated with mLOY. However, except smoking, all the other traits failed to reach the significance level after Bonferroni correction. There were several traits that reached the nominal significance level (*p*-value<0.05). Serum creatine and uric acid levels showed evidence of association with mLRRY, with 0.04 (95%CI, 0.01-0.07) and 0.03 (95%CI, 0-0.07) SD increase of the mLRRY values per 1-SD increase of the two risk factors, respectively, suggesting that reduced renal function and hypouricemia may affect mLOY (Figure 5.4). Increased mLRRY were also associated with lower levels of transferrin, with 1-SD lowering of transferrin levels reducing mLRRY by 0.04 SD (95%CI, 0-0.07SD, Figure 5.4). Lower levels of transferrin, pathologically ascribed to impaired production of transferrin in the liver or excessive loss through the kidney, suggested impaired capacity of iron transportation into circulation, leading to iron deficiency anaemia. Moreover, reduced mLRRY was associated with higher levels of glycated haemoglobin (HbA1c), a biomarker routinely measured for diagnosis and monitoring of T2D, with 1-SD increase of HbA1c levels reducing mLRRY by 0.04 SD (95%CI, 0-0.07SD, Figure 5.4).

Figure 5.4. Associations of mLRRY with lifestyle and anthropometry traits and circulatory biomarkers in the random subcohort of EPIC-InterAct study.

The unit of X-axis is SD change of normalised mLRRY per one unit (SD or categorical level) increase of the main exposure.



5.4 Discussion

5.4.1 Summary and conclusion

The main finding of this study is that there was some suggestive evidence showing mLRRY, the measure of mLOY, was associated with increased risk of T2D in the model adjusted for centre, array and age, but the effect was weak, and substantially reduced after adjusting for smoking. In addition, in UK biobank, the association was observed in an opposite direction, causing the findings to become even more inconclusive.

Moreover, in EPIC-InterAct, the risk effects of mLRRY on incident T2D, although non-significant, were proportionally increased to the increasing order of quartiles of mLRRY, suggesting a linear trend in associations between mLRRY and T2D risk, even though the overall effect of mLRRY on T2D risk was weak. Men younger than 50 years were more susceptible to T2D with decreased mLRRY than men older than that age. Finally, some modifiable risk factors showed nominally significant associations with mLRRY, yet all below Bonferroni corrected threshold, except smoking status.

5.4.2 Age-stratified risk effects of mLOY on T2D

The risk effect of mLOY on T2D was stronger in the young (< 50 years of age) than the middle-aged (50-65 years of age) or old (> 65 years of ages) populations. A statistically significant interaction between mLOY and age group was observed for their risk effects on T2D. Post-zygotic mutations occur stochastically across human genomes, among which those that confer proliferative advantages can lead to aberrant clonal expansions[200]. Such mosaic mutations are often rarely observed in young individuals, although they have also been described in new-born babies with significantly higher risk of leukaemia[200,382], suggesting that they can occur throughout the entire human lifespan, but are enriched in the elderly. In our study, we found that higher degrees of mLOY (mLRRY<-0.15) began to emerge after the age of 50 years, hence in the younger population, whether the mLRRY values indicate true mLOY events or merely reflect some background noises is unclear. Therefore, although we observed a larger risk effect of mLOY on T2D in the younger population, we cannot rule out the

possibility that this was induced by technical errors.

5.4.3 mLOY mediating the risk effect of smoking on T2D

Previous studies have suggested a causal effect of smoking on mLOY under the MR framework, where genetic variants at the *CHRNA5–CHRNA3–CHRNB4 nicotinic receptor* locus were used as instrumental variables[195,196]. Given that smoking has also been reported as a modifiable risk factor for T2D independent of educational level, physical activity, alcohol consumption, and diet[381], and adjusting for smoking substantially attenuated the risk effect of mLRRY on T2D, we speculated that mLOY may be one of the pathological consequences of smoking, thereby partially mediating the risk effect of smoking on T2D. In other words, mLOY may act as a mediator that contributes to the detrimental effect of smoking on T2D.

5.4.4 Discordant findings between EPIC-InterAct and UK Biobank

We showed that in UK biobank, the association between mLRRY and incident T2D became much stronger, but completely reversed. There are several potential factors that might partially explain the discrepant findings between EPIC-InterAct and UK biobank. Firstly, the study design and participant demographics can differ; EPIC-InterAct is a case-cohort study with enriched T2D cases that account for over half of the overall sample size, whereas UK biobank is a population-based cohort study, prospective in design, but so far with limited longitudinal data. The proportion of incident T2D cases is relatively small in UK biobank, accounting for approximately 4% of the study population, and therefore leading to an unbalanced case-control ratio that can result in type I error rate inflation due to violation of asymptotic assumptions of logistic regression[383]. Secondly, definition of incident T2D between two studies can differ. Because the aim of the EPIC-InterAct study is T2D-focused, the T2D cases have been ascertained through comprehensive reviewing of various sources of evidence, including questionnaire and nurse interview, linkage to primary and secondary care registers, medication use, and mortality data[212]; whereas in UK biobank, the incident T2D cases were defined primarily based on electronic health records data through linkage to ICD10 codes. The less stringent ascertainment of incident T2D in UK biobank may cause

inaccurate or even biased estimation of associations. Thirdly, the model diagnostic statistics suggested a potential multi-collinearity problem in UK biobank, and thus a risk of global instability of model estimates. Moreover, the intercept terms from logistic and Cox regression models were highly correlated to age and centre, which also indicated instability of the models applied in UK biobank. Because of the massive and complex data structure in UK biobank, there might be unknown confounding factors that are essential but underexplored, such that omission of them can lead to biased results. Investigations into the model reliability and better refined algorithms for defining T2D incidence in UK biobank may help to address such discrepancies and improve research reproducibility.

5.4.5 Impact and strength

Post-zygotic variations sporadically emerged in somatic cell lineages from the zygote stage onwards throughout the entire lifespan are largely under-investigated[200], and they have been demonstrated to underlie pathophysiology of various clinical disorders, including early developmental defects and age-related diseases, such as cancer predisposition (both within and outside the haematological system), cardio-metabolic phenotypes and AD[189,190,207]. A recent study collecting data from several case-control cohorts (n=7,437, 29.7%) has shown T2D increased the average risk of clonal mosaic events in autosomes by more than 5-fold[360], but whether clonal mosaicism, and particularly on chromosome Y, affects T2D risk is unknown. This study, with an aim to answer this question, is the first study that has examined observational association of mLOY with T2D in a large prospective cohort. It benefited from a large-scale case-cohort study design, comprising a large sample size of incident T2D cases with diagnoses verified during a long-term follow-up, as well as a quasi-random subcohort. A variety of phenotypes were measured, which allowed comprehensive adjustments for potential confounding factors.

5.4.6 Limitation and future perspectives

The analyses showed a lack of statistical power, especially when dichotomising mLRRY or categorising it into quartiles. Substantial between-country heterogeneities were observed, which also compromised the statistical power. Moreover, the observational associations

analysed in this study can be biased by confounding factors and reverse causation, while causal evidence is lacking. MR analyses with strong and specific genetic instruments for mLOY can help to further dissect potential causal roles of mLOY on T2D. Given that little evidence has been found to suggest an observational association, to demonstrate a convincing causal association, MR studies need to be well-sufficiently powered, and with genetic instruments specifically capturing phenotypic variations of mLOY. Finally, this study analysed clonal mosaic events on chromosome Y only, given that previous studies have suggested a strong correlation between mLOY and clonal mosaic events across autosomal chromosomes[206], and between mLOY and mLOX in women[195], a further investigation of associations of overall and site-specific clonal mosaicism with T2D risk would be of interest.

Chapter 6

Summary and discussion

6.1 Key findings

Several key findings have arisen from the work presented in my thesis. (1) I identify 49 genomic regions associated with LTL, and prioritise likely-causal gene candidates at 32, many of which are related to telomere structure, DNA replication and repair. Nucleotide metabolism is highlighted as an important regulatory pathway of TL for the first time, using a data-driven method in combination with knowledge-based manual curation. (2) Using genetic approaches to understand causal associations between LTL and disease, I report several associations not previously identified, including a higher risk of hypothyroidism and lower risks of thyroid cancer, lymphoma and several malignant as well as benign tumours, particularly of proliferative tissues. I replicate the previously reported association of shorter LTL with increased risk of CHD but show that this association appears to be mainly driven by specific genes that exert relatively large and directionally concordant effects on CHD. (3) I suggest that investigating longitudinal changes of LTL in a large prospective cohort can be theoretically valuable, but practically challenging. My systematic literature review summarises epidemiological studies of risk factors that are associated with the longitudinal changes of LTL, i.e. attrition rate, and how such changes are consequently associated with disease outcomes. Results show that evidence is inconsistent, except for baseline LTL being a strong determinant of the LTL attrition rate, the association that I also observe in the small pilot study within the Fenland cohort. Fenland results in combination with evidence from the literature review question the feasibility and scientific efficiency of studies of longitudinal change of LTL, given the strong correlation of baseline LTL with the prospective 'rate of change'. (4) I find little evidence to suggest an observational association between mLOY and T2D risk. Although there is some suggestive evidence for a modest effect of mLOY on the risk of T2D in EPIC-InterAct case-cohort study, the effect is largely attenuated to be non-significant after adjusting for smoking. In addition, such association appears to be in an opposite

direction in UK biobank. Therefore, considering confounding effects and inconsistency across study cohorts, I conclude that there is insufficient evidence to suggest mLOY is observationally associated with T2D risk.

This chapter summarises findings and conclusions from the work undertaken across all chapters in this thesis, with a focus on the overall impact, strengths and limitations of my work. Future directions for studying genomic markers of aging and their potential utilities in clinical healthcare and drug development are highlighted in the last part of this chapter.

6.2 Summary and discussion

This thesis focuses on genomic ageing in the population and includes detailed studies of key markers of genomic ageing: LTL and mLOY. Original research has been performed to characterise novel genetic determinants and biological mechanisms of LTL regulation and dissect causal roles of LTL in various clinical outcomes. Moreover, literature evidence on risk factors that are associated with longitudinal changes of LTL and clinical consequences due to such changes has been systematically reviewed. Feasibility of studying LTL shortening in a young and healthy cohort has been examined within a small-scale pilot study in Fenland. Finally, the observational association between mLOY and incident T2D risk has been characterised in two of the largest international prospective studies of incident T2D, EPIC-InterAct and UK biobank.

Strengths and limitations specific to each study have been discussed in each relevant chapter and this section focuses on some of the more general aspects of the work undertaken in my thesis. The GWAS meta-analysis of LTL substantially increases sample sizes of previous studies by adding two large-scale population cohorts, EPIC-InterAct and EPIC-CVD, and SNP coverage has been increased via upgraded, more densely imputed panels. Overall, this study has more power to discover rare and low frequency variants that were difficult to study in the past due to a lack of statistical power and/or accurate genotyping/imputation of the variants. Moreover, with rich sources of publicly available data for variant and gene level functional annotations, and cutting-edge analytical methods of integrating such data with genomic results, I generate more evidence to help with pinpointing likely-causal genes for associated LTL loci identified. Through analysing functional inter-connections between likely-causal

genes, I prioritise signalling pathways, some of which have not been previously implicated with telomere biology. Furthermore, owing to increasingly expanding international consortia that provide GWAS summary statistics, I dissect causal roles of LTL in cardiometabolic traits and diseases with substantially increased power. I also systematically elucidate clinical relevance of LTL in a broad spectrum of disease outcomes via a phenome-wide association scan in UK biobank; many of the diseases have not been investigated before.

6.2.1 Genetic architecture of LTL

Despite doubling sample size of the recent genome-wide meta-analysis of LTL[151], I identify only a relatively small number of loci not previously identified, and this can be attributed to several reasons. (a) A couple of GWAS of LTL, similar to this work, have been published very recently, and some of their novel loci overlap with ours, which are not counted as novel in our study. However, our work is independent from theirs, and has been written up and was under internal review by the time these publications emerged. These include a multi-ethnic analysis in TOPMed (Trans-Omics for Precision Medicine) (n=75,000) with LTL estimated using whole genome sequencing data[384], and a meta-analysis of results from a Singaporean Chinese study and the previously published ENGAGE study (n=60,601)[240]. (b) Rare variant discovery: Firstly, given that extremely shortened LTL is often observed in patients with premature ageing syndromes, who die at fairly young ages, mutations of these disease-causing genes, if at a relatively high level of penetrance, are unlikely to be transmitted to subsequent generations. Therefore, even though their effect sizes are large, their observed frequencies in a general population are too low to be genotyped or imputed accurately. Secondly, even if they can be accurately genotyped, for example, through direct genotyping or whole genome/exome sequencing, identification of them suffers from a lack of statistical power, and finding such associations needs much larger sample sizes. (c) identification of common variants with small effects: Despite conducting the largest existing GWAS on LTL, genome-wide independent variants identified explain only 1.5% of the total heritability. The missing heritability is likely to be explained by both unidentified rare variants with large effects as well as many more common variants with small effects that this study was still not powered to detect. In comparison to other complex traits, such as lipids[385,386] and blood pressure[387], where hundreds of independent variants that span a wide range of allele frequencies have been identified in recent multi-ethnic GWAS meta-analyses in large-scale biobanks, genetic

discovery on LTL is restricted by limited samples available possibly due to measurement efficiency and accuracy. Future meta-analyses of studies from biobanks and consortia can help to expand the catalogue of genetic determinants of LTL to include ultra-rare variants with extremely large effect sizes, as well as common variants with very small effects[388].

6.2.2 Causal gene annotation

Identifying causal variants and genes underlying association signals has been a general challenge in genetic association studies. To address this, I gathered and integrated extensive genomic and regulatory annotations and gene expression data, in addition to manual curation with help and guidance from senior study authors of the related publication. Although these efforts have successfully resulted in over two thirds of the FDR loci being assigned to likely-causal gene candidates, annotation was difficult for the remainder due to inconsistent prediction results of likely-causal genes by different methods and/or limited knowledge of candidate genes within those loci. Large-scale GWAS on other traits typically rely on one or two methods of causal gene prioritisation, and therefore do not have such problem of showing inconsistent results across methods used. Different algorithms implemented and training data used may lead to discrepancies observed within the prediction results of causal gene candidates, however, no studies have systematically compared and evaluated strengths and limitations of these methods. Therefore, I employed a conservative approach, where only genes supported by multiple bioinformatic evidence showing consistent results are deemed as likely-causal genes. Given that rare variants with extreme effects on protein functions can facilitate causal gene prioritisation, exome and whole-genome sequencing at scale, as planned and ongoing for UKBB and other studies, can help to pinpoint likely-causal genes for these loci and those identified in the future.

6.2.3 Trans-ethnic analyses

This and other recent LTL GWAS are predominantly Europeans-focused, with the exception of the Singaporean Chinese study (n=23,096 with Southern Han Chinese ancestry)[240] and the TOPMed study (n=75,176 with 28% African, 13% Hispanic/Latino, 6% Asian and 2% Samoan ancestries)[384]. Under-representation of non-European populations is a recognised and important limitation of most genetic population studies. Inclusion and study of diverse

ethnicities are valuable for several reasons: First, novel loci can be identified through meta-analysing studies across populations. Because variants may show distinct allele frequencies in different populations, loci that are monomorphic or rare in the European populations can be polymorphic and common in other populations, therefore meta-analyses can increase the power of identifying rare variants-driven loci that may be missed in studies with European ancestry samples only. Second, multiple conditionally independent variants can be identified with improved fine-mapping resolution by leveraging ethnic differences in LD structure. For example, studies have shown that inclusion of African ancestry samples leads to marked improvement in localisation of causal variants because of low LD and high genetic heterogeneity within the African genomes[389]. Third, generalisability of scientific and clinical utilities of summary statistics from GWAS can be improved. For example, the optimal choice of SNPs and weights for PRS construction may differ between populations due to different LD and allele frequency patterns, increasing diversity of sample ethnicities in GWAS can help improve prediction accuracy of PRS in individuals of non-European ancestries[81]. Although some efforts have shown promise in levelling the imbalance of sample ethnicities, in general, cross-ethnic analyses are still dominated by Europeans. Future research that includes larger proportions of individuals with more diverse ancestries may further increase the power of identifying novel loci and variants involved in LTL regulation, and facilitate biomedical applications of GWAS results in wider populations.

6.2.4 Measurement of LTL and longitudinal assessment

Imprecise measurement of LTL is another potential limitation of this study. We used the qPCR method for LTL measurement, which may be less reliable than other procedures like Southern Blot, but more suitable for large-scale studies due to less amounts of DNA samples required and time and cost efficiency. However, there are several issues of the qPCR measurement, which can result in power loss and false positive discoveries. First, a relatively large random variation within cross-sectional measures reduces statistical power in regression analyses. To control such random errors, we measured all samples in triplicate and excluded samples with coefficient of variation larger than 10%. Second, batch- and centre-effects, i.e. heterogeneities of measurement between different batches and centres due to different experiment times, handlers, reagent lots etc., decrease power of meta-analyses while increase false positive findings. To control these, we adjusted for batch and centre in linear

regression models and applied *post-hoc* corrections in meta-analyses to filter out variants with substantial between-study heterogeneities.

In terms of repeated LTL measures, i.e. longitudinal change of LTL, random errors within repeated measures of same individuals over time can mask real longitudinal changes due to the regression to the mean problem (section 4.4.2). Such problem can lead to an artificial correlation between the change (follow-up minus baseline measurements) and the baseline of measurements, because individuals whose baseline measures are lower than average tend to increase (so that change values are larger) and vice versa. This is specifically true for genetic discovery studies, where small effect sizes are expected and for which very large-scale (cross-sectional) GWAS now exist on LTL and hence new longitudinal studies are likely to only identify established loci already known to affect baseline LTL. This technical challenge questions the feasibility of conducting large-scale epidemiological and GWAS on the longitudinal change of LTL.

While protocol prioritisation of the existing method and extensive QC procedures have been undertaken, alternative methods may be worth trying in future researches, such as estimation of LTL from whole genome sequencing data[390]. The recent TOPMed study has leveraged computational methods using the whole genome sequencing data, which made a marked progress in realising high-throughput and fully unsupervised measurement of LTL, however, showed only a moderate level of correlation (Pearson correlation $r<0.60$) to Southern blot estimation[396], highlighting the challenge and technical enhancement required for further improvement of accuracy and efficiency of the LTL measurement.

6.2.5 Measurement of mLOY

Population studies on clonal mosaicism, including mLOY, have largely been limited by technical challenges of accurately detecting such events, due to expected low cell fractions of mosaic events in leukocytes. In this study, a mLRRY approach that quantifies mLOY based on median genotyping intensity over the Y chromosome specific (non-pseudoautosomal) region was used. This method has been previously used in a GWAS meta-analysis of mLOY and showed potentials of being implemented to large-scale cohort studies with genotyping array data available. However, as this method needs to be applied to different array data separately, heterogeneity between arrays can be an issue that decreases power when meta-analysing across arrays. Moreover, such estimation was found to be missing in up to 50% of men in the

EPIC-InterAct study, and the missingness was shown to be correlated to genotyping array, casting further doubts on the accuracy and stability of this method of measurement.

Recently, a more refined approach has been developed, which uses signal intensity imbalance between two statistically phased haplotypes over the pseudo-autosomal region of the X and Y chromosomes[135]. It showed improved precision of estimation compared to the mLRRY approach, which, if applied to the EPIC-InterAct study, can potentially increase the power of observational analyses. This is especially important considering the observed association between mLRRY and incident T2D, if present, is weak, and suffers from a lack of statistical power.

6.3 Future work and applications

Telomeres

Restricting telomere elongation has been proposed as a tumour suppression mechanism but shortened telomeres can in turn influence stem cell differentiation and tissue renewal, highlighting an important role of telomere homeostasis in various diseases, including cardio-metabolic and neurodegenerative diseases, cancers, as well as rare diseases of premature ageing (section 1.4.1.4).

6.3.1 TL and premature ageing syndromes

My work and previous studies have highlighted overlaps between genetic determinants of premature ageing syndromes and normal ageing-related phenotypes and diseases. Therefore a deeper understanding of mechanistic pathways underlying the pathogenesis of rare and extreme ageing syndromes, such as HGPS and WS, can help to also shed light upon pathophysiological changes that occur during ageing age-related diseases in general populations and may provide novel therapeutic approaches for treating both rare premature ageing syndromes and common age-related complex diseases.

One example for an overlapping genetic mechanism between premature ageing syndromes and telomere regulation is via the *WRN* gene, the causal gene for WS, which has a well-established role in homology-dependent recombinational DNA repair and telomere maintenance. Loss of *WRN* leads to critically shortened telomeres and genomic instability[124], the hallmarks of ageing (Figure 1.1). Another example is an alternative splicing mutation of a

LMNA gene transcript, the most frequent cause of the HGPS, resulting in protein truncation from the C terminus of lamin A[123]. The truncated protein mutant is called progerin and several studies have indicated that progerin modifies nuclear environment in which DNA repair pathways are activated[125]. Cell samples from HGPS patients show reduced recruitment of protein components of DDR pathway at DNA double-stranded breaks, leading to impaired genome integrity and cell proliferative capacity[122,123]. In chapter 2, I identify multiple genes (*ATM*, *PARP1*, *TERF2*, *SENP7* and *RFWD3*) that are associated with LTL and involved in the DDR pathway (section 2.3.3). Mutations in these genes have been shown to disrupt telomere homeostasis and lead to premature ageing phenotypes[172–174], suggesting DDR as a potential mechanistic pathway that conveys genetic effects of telomere dysregulation onto organismal phenotypes that resemble ageing. It remains uncertain whether TL shortening lies on the causal path that leads to clinical phenotypes manifested in the premature ageing syndromes or merely a consequence of shared genetic factors that drive both TL shortening and accelerated ageing phenotypes. Genetic and molecular characterisation of TL can deepen our understanding of mechanisms of rare disorders of premature ageing, thereby providing an aetiological link between premature ageing syndromes and common complex diseases occurring during normal ageing, and eventually facilitating novel therapeutic target discovery and prevention and treatment of age-associated diseases.

6.3.2 TL and age-related complex diseases

Besides rare disorders of premature ageing, TL has been causally linked to various complex morbidities that occur more frequently in general populations, including CVD[158,177,291], AD[180], dementia and mortality[181]. The causalities are mainly inferred from genomic research where genetic determinants of TL are used as instruments within the MR framework (section 1.2.6.1). However, mechanistic pathways of how telomeres are involved in age-related complex diseases, for example, which genes regulating TL influence disease risks are poorly understood. These are important questions for translating TL into clinical applications and for developing cancer immunotherapies that target TL-associated genes[391].

6.3.2.1 TL and CVD

Endothelial cell senescence triggered by critically shortened TL at atherosclerotic lesions has been shown to contribute to atherogenesis, providing a mechanistic link between shortened TL and increased CVD risk[392]. Although TL exhibits a certain level of heterogeneity across tissues, for example, LTL was reported to be shorter than TL in vascular tissues[393], the high correlation between tissues and the feasibility of measuring LTL at large-scale make LTL an excellent proxy for TL across tissues[392]. However, several prospective population-based studies casted doubts on the concept that longer TL can always protect against CVD, depending on disease onset and progression[394–396]. Moreover, additional studies that showed supporting evidence for the protective effects of longer TL on CVD are often debatable due to relatively small effect sizes reported. Alternative hypotheses have been postulated that suggest other aetiological causes of CVD, including oxidative stress and inflammation, influence LTL attrition rate[397], which in turn result in the observed association between cross-sectional measurement of LTL and CVD. Such caveat of reverse causation has been discussed in relevant sections (sections 1.4.1.4.2 and 3.4.2).

6.3.2.2 TL shortening and CVD

Studies have suggested that genetic determinants of LTL shortening rate may be more directly relevant to CVD aetiologies compared to those of LTL measures at birth[391]. Identification of genetic determinants of LTL shortening rate have presumably been included in GWAS of cross-sectional measurement of LTL, because such GWAS should identify genetic factors that regulate LTL at two levels: LTL at birth and LTL shortening rate. For instance, genes that encode protein components of the core telomere structure and the telomerase ribonucleoprotein, can constitutively regulate TL from birth and throughout the entire life course. Therefore, these genes can determine TL at birth as well as TL attrition later in life. In contrast, genes that function in inflammatory pathways, such as HLA-mediated immune responses and stress-triggered DDR, may regulate TL attrition rate via interactions with physiological stress conditions, such as oxidative stress and nutrient deficiency.

Age-related LTL attrition rate seems to reach the highest level early in life, and then sustain at a relatively low level throughout the adulthood, suggesting that interindividual variation of LTL is mainly dependent on LTL at birth and attrition rate during childhood rather than middle or late adulthood when CVD risk assessment is warranted[398]. A study cohort of

adult participants should capture the main sources of pre-disease variation of LTL with only baseline measurement. Moreover, this and earlier work has shown that LTL shortening rate is highly correlated with baseline LTL (Chapter 4). Because of these, given the fact that a simple, single point measure a) reflects genetic influences as well as accumulative burdens of non-genetic environmental risk effects on LTL, and b) is a strong determinant of LTL shortening rate, it is unlikely that assessing LTL shortening rate can provide any additional values in either identifying genetic determinants associated with LTL or predicting risk of CVD.

6.3.2.3 Gene-specific effects of LTL on CVD

Because candidate genes identified in GWAS of LTL can have very different biological functions, and thus may be directly or indirectly associated with telomere biology. Their individual involvements in telomere regulation are likely to vary and their impact on CVD risk may be mediated through distinct mechanisms rather than LTL alone. For example, telomerase was found to be involved in the Wnt/β-catenin pathway and the E2F1-based transcription, both of which have been implicated in atherosclerosis independently of TL[399,400]. Therefore, although genetically determined shorter LTL increases the risk of CVD, potential pleotropic effects of variants used to proxy LTL may undermine this conclusion, and further considerations about individual gene-specific effects may be necessary.

6.3.2.4 Conclusion

The telomere hypothesis is attractive in that rather than a single gene test, it provides a global property of the genome that is both heritable itself and modifiable by other conventional risk factors of CVD, thereby serving as an integrative marker of biological age that could at least partially explain interindividual variation in risk of occurrence and age of onset[94]. However, given the methodological challenges of assessing LTL attrition rate simply and accurately, and the relatively small changes that are observed during middle age, distinguishing individuals' CVD risks based on repeated measures of LTL (i.e. longitudinal change) seems unfeasible at this stage for the reasons mentioned above[391].

mLOY

mLOY in peripheral blood is reportedly the most common form of clonal mosaicism (post-zygotic mutations) in men during a physiological process of ageing, with a prevalence of up to 20% (among three prospective cohorts that include 8,679 cancer cases and 5,110 cancer-free controls) in men over 80-years[196]. Besides age, smoking is a well-documented risk factor for mLOY, with more than 3-fold increase of mLOY risk in current versus never smokers[401], although this effect was reported to be transient, as cessation of smoking for several years reduced the risk towards baseline[196]. Several recent studies have linked mLOY in blood leukocytes to higher risks of all-cause mortality and common age-related diseases, including AD,[207] severe atherosclerosis[356], and cancers, not only in blood but many other tissues[193,200,359], but the extent to which these observations represent causations is unclear. Understanding fundamental mechanisms of how mLOY in leukocytes affect occurrence and progression of age-related diseases is important for assessing clinical significance of mLOY as a biomarker for these diseases[135,402].

6.3.5 mLOY and T2D

6.3.5.1 Observational association between mLOY and T2D

In Chapter 5, I have demonstrated that mLOY exerts little effect on incident T2D in a large European prospective case-cohort study (EPIC-InterAct). There is no association after adjusting for important confounding factors, including age and smoking. However, the association between mLOY and T2D risk appeared somewhat more pronounced in younger compared to older men. Whether mLOY in younger men is truly involved in T2D pathology or acts as a molecular mediator of conventional risk factors of T2D not accounted for in the present study is unknown. Also, mLOY may simply serve as a marker of genome instability that underpins various age-related disorders, but not directly contribute to disease aetiologies. To further elucidate the role of mLOY in T2D aetiology, further investigation into common genetic mechanisms that drive mLOY and susceptibility to T2D would be crucial.

6.3.5.2 Genetic determinants of mLOY and T2D

A recently published GWAS identified 19 genetic loci that are associated with mLOY, many of which are functionally implicated in cell cycle regulation and DNA repair[195]. These help to highlight relevance of mLOY in carcinogenesis due to substantial overlap of mLOY-associated loci with known oncogenic genes and gene targets for cancer therapies. A larger GWAS with

improved methods of detection of mLOY is underway, providing more novel loci and deeper insights into their clinical relevance[135].

Several mLOY-associated loci identified in the ongoing GWAS of mLOY are correlated with loci previously reported for T2D risk, highlighting several genes that encode for cyclins and cyclin-dependent kinases expressed in pancreatic β cells, essential for β-cell growth and maturation[135]. These may reflect plausible involvement of cell cycle regulation and genome integrity in β-cell proliferation and differentiation, and thus linking mLOY to insulin-dependent diabetic phenotypes. However, because such relevance is driven by specific genes, and thus does not reflect a causation of mLOY on T2D, but merely suggests a common genetic susceptibility to genome instability that leads to both mLOY and β-cell loss. Moreover, reduced β-cell mass is a pathological feature of T1D rather than T2D, the latter is caused by insulin resistance in peripheral tissues, but not reduced insulin secretion from β cells[360]. Therefore, although there are some genetic evidence linking mLOY to metabolic disorders, these do not counterpose our conclusion in the prospective cohort study mentioned above.

6.3.5.3 Conclusion

Our epidemiological analyses suggest no association between mLOY and incident T2D, and with genetic discovery from previously published work, these can help to facilitate our understanding of the role of mLOY in the risk and aetiology of T2D.

6.4 Conclusions

The work conducted in this thesis brings together evidence from large-scale studies of two genomic markers of ageing, LTL and mLOY, including assessments of genetic and non-genetic determinants and clinical consequences, with a focus on T2D. The GWAS meta-analysis of LTL with downstream *in silico* annotations provides an expanded pool of likely causal genes implicated in telomere homeostasis and highlights novel biological mechanisms regulating LTL for experimental follow-up. LTL attrition is observed in a relatively young and healthy contemporaneous population but studying it at scale is questioned with technical and statistical challenges. I demonstrate a fundamental implication of LTL in cancer development, whereas no association is observed with T2D for either of the two markers of genomic ageing.

Overall, this work helps to build the foundations for future studies in exploring causal roles of genomic ageing markers in various age-related diseases, as well as investigating their potential values in predicting risks of these diseases.

References

1. Singh, P. P., Demmitt, B. A., Nath, R. D. & Brunet, A. The Genetics of Aging: A Vertebrate Perspective. *Cell* **177**, 200–220 (2019).
2. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, (2013).
3. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
4. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
5. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, (2017).
6. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).
7. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).
8. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* **291**, 1304 LP – 1351 (2001).
9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **304**, 1321 LP – 1325 (2004).
11. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
12. Altshuler, D., Donnelly, P. & Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
14. Consortium, the H. R. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279 (2016).

15. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
16. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–50 (2014).
17. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).
18. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
19. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
20. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
21. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
22. Wray, N. R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res. Hum. Genet.* **8**, 87–94 (2005).
23. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
24. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).
25. Chen, B., Cole, J. W. & Grond-Ginsbach, C. Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Front. Genet.* **8**, 167 (2017).
26. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
27. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
28. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
29. Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 567–573 (2008).
30. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide

association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).

31. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
32. Li, M.-X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
33. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
34. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
35. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
36. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, 434–42 (2004).
37. Wakefield, J. A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
38. DeWan, A. *et al.* HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration. *Science (80-.).* **314**, 989 LP – 992 (2006).
39. Consortium, T. W. T. C. C. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).
40. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, (2018).
41. Marigorta, U. M. & Navarro, A. High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLOS Genet.* **9**, e1003566 (2013).
42. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255 (2008).
43. Mackay, T. F. The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**, 303–339 (2001).
44. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–7 (2016).

45. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17**, 782 (2014).
46. Polychronakos, C. & Li, Q. Understanding type 1 diabetes through genetics: advances and prospects. *Nat. Rev. Genet.* **12**, 781–792 (2011).
47. Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLoS Genet.* **7**, e1002293 (2011).
48. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110 (2017).
49. Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun.* **9**, 260 (2018).
50. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet (London, England)* **376**, 180–8 (2010).
51. Wu, M. C. *et al.* Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am. J. Hum. Genet.* **86**, 929–942 (2010).
52. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
53. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *bioRxiv* (2019).
54. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
55. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
56. Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). The ENCODE Project Consortium. *PLoS Biol.* **9**, e1001046 (2011).
57. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
58. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
59. GTEx consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
60. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using

reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

61. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–52 (2016).
62. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **0**, 473–487 (2017).
63. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
64. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
65. Bothwell, L. E., Greene, J. A., Podolsky, S. H. & Jones, D. S. Assessing the Gold Standard — Lessons from the History of RCTs. *N. Engl. J. Med.* **374**, 2175–2181 (2016).
66. Altman, N. & Krzywinski, M. Association, correlation and causation. *Nat. Methods* **12**, 899–900 (2015).
67. Davies, N. M., Holmes, M. V & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
68. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
69. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Stat. Med.* **35**, 1880–1906 (2016).
70. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).
71. Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
72. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).

73. Swerdlow, D. I. *et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45**, 1600–1616 (2016).
74. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–14 (2016).
75. Maas, P. *et al.* Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol.* **2**, 1295 (2016).
76. Desikan, R. S. *et al.* Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLOS Med.* **14**, e1002258 (2017).
77. Seibert, T. M. *et al.* Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, j5757 (2018).
78. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* **375**, 2349–2358 (2016).
79. Ibanez, L., Farias, F. H. G., Dube, U., Mihindukulasuriya, K. A. & Harari, O. Polygenic Risk Scores in Neurodegenerative Diseases: a Review. *Curr. Genet. Med. Rep.* **7**, 22–29 (2019).
80. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
81. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
82. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013).
83. Richardson, T. G., Harrison, S., Hemani, G. & Davey Smith, G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* **8**, (2019).
84. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
85. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
86. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-

mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).

87. McCarthy, M. I. & Mahajan, A. The value of genetic risk scores in precision medicine for diabetes. *Expert Rev. Precis. Med. Drug Dev.* **3**, 279–281 (2018).
88. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
89. Meigs, J. B. *et al.* Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).
90. Lango, H. *et al.* Assessing the Combined Impact of 18 Common Genetic Variants of Modest Effect Sizes on Type 2 Diabetes Risk. *Diabetes* **57**, 3129–3135 (2008).
91. van Hoek, M. *et al.* Predicting Type 2 Diabetes Based on Polymorphisms From Genome-Wide Association Studies: A Population-Based Study. *Diabetes* **57**, 3122–3128 (2008).
92. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
93. World Health Organization. World Report on Ageing and Health. *www.who.int* (2015).
94. Samani, N. J. & van der Harst, P. Biological ageing and cardiovascular disease. *Heart* **94**, 537–9 (2008).
95. Finkel, T., Serrano, M. & Blasco, M. A. The common biology of cancer and ageing. *Nature* **448**, 767–774 (2007).
96. Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNf676 as genes regulating telomere homeostasis in humans. *Hum. Mol. Genet.* **21**, 5385–5394 (2012).
97. McDaid, A. F. *et al.* Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.* **8**, 15842 (2017).
98. Graham Ruby, J. *et al.* Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* **210**, 1109–1124 (2018).
99. Partridge, L. & Gems, D. Mechanisms of ageing: public or private? *Nat. Rev. Genet.* **3**, 165–75 (2002).
100. Christensen, K., Johnson, T. E. & Vaupel, J. W. The quest for genetic determinants of human longevity: challenges and insights. *Nat. Rev. Genet.* **7**, 436–448 (2006).
101. Kenyon, C. J. The genetics of ageing. *Nature* **464**, 504–512 (2010).
102. Johnson, S. C., Rabinovitch, P. S. & Kaeberlein, M. mTOR is a key modulator of ageing

and age-related disease. *Nature* **493**, 338–345 (2013).

103. Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
104. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
105. Kimura, K. D., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. daf-2, an Insulin Receptor-Like Gene That Regulates Longevity and Diapause in Caenorhabditis elegans. *Science (80-.).* **277**, 942–946 (1997).
106. Milman, S. *et al.* Low insulin-like growth factor-1 level predicts survival in humans with exceptional longevity. *Aging Cell* **13**, 769–771 (2014).
107. Suh, Y. *et al.* Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc. Natl. Acad. Sci.* **105**, 3438–3442 (2008).
108. Martins, R., Lithgow, G. J. & Link, W. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell* **15**, 196–207 (2016).
109. Flachsbart, F. *et al.* Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2700–5 (2009).
110. Youngman, L. *et al.* Protein oxidation associated with aging is reduced by dietary restriction of protein or calories. *PNAS* **89**, 9112–9116 (2008).
111. Broer, L. *et al.* GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **70**, 110–118 (2015).
112. Flachsbart, F. *et al.* Identification and characterization of two functional variants in the human longevity gene FOXO3. *Nat. Commun.* **8**, 2063 (2017).
113. Sun, L. Y. *et al.* Longevity is impacted by growth hormone action during early postnatal period. *Elife* **6**, e24059 (2017).
114. Ben-Avraham, D. *et al.* The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with increased GH sensitivity and taller stature. *Sci. Adv.* **3**, e1602025 (2017).
115. van den Berg, N., Beekman, M., Smith, K. R., Janssens, A. & Slagboom, P. E. Historical demography and longevity genetics: Back to the future. *Ageing Res. Rev.* **38**, 28–39 (2017).
116. Giuliani, C., Garagnani, P. & Franceschi, C. Genetics of Human Longevity Within an Eco-Evolutionary Nature-Nurture Framework. *Circ. Res.* **123**, 745–772 (2018).

117. Albani, D. *et al.* Modulation of human longevity by SIRT3 single nucleotide polymorphisms in the prospective study “Treviso Longeva (TRELONG)”. *Age (Omaha).* **36**, 469–478 (2014).
118. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* **8**, 1–40 (2019).
119. Joshi, P. K. *et al.* Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nat. Commun.* **8**, 1–13 (2017).
120. Joshi, P. K. *et al.* Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nat. Commun.* **7**, 11174 (2016).
121. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing. *Nature* **561**, 45–56 (2018).
122. Scaffidi, P. & Misteli, T. Lamin A-dependent nuclear defects in human aging. *Science* **312**, 1059–63 (2006).
123. De Sandre-Giovannoli, A. *et al.* Lamin a truncation in Hutchinson-Gilford progeria. *Science* **300**, 2055 (2003).
124. Zhang, W. *et al.* Aging stem cells. A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science* **348**, 1160–3 (2015).
125. Kudlow, B. A., Kennedy, B. K. & Jr, R. J. M. Werner and Hutchinson – Gilford progeria syndromes : mechanistic basis of human progeroid diseases. *Nat. Rev. Mol. Cell Biol.* **8**, 394–404 (2007).
126. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23**, 4420–4432 (2014).
127. Newman, A. B. *et al.* A Meta-analysis of four genome-wide association studies of survival to age 90 years or older: The cohorts for heart and aging research in genomic epidemiology consortium. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **65 A**, 478–487 (2010).
128. Broer, L. *et al.* Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur. J. Hum. Genet.* **21**, 1163–1168 (2013).
129. Walter, S. *et al.* A genome-wide association study of aging. *Neurobiol. Aging* **32**, (2011).

130. Pilling, L. C., Atkins, J. L., Bowman, K. & Jones, S. E. longevity is influenced by many genetic variants : evidence from 75 , 000 UK Biobank participants. *Aging (Albany. NY).* **8**, 547–560 (2016).
131. Fortney, K. *et al.* Genome-Wide Scan Informed by Age-Related Disease Identifies Loci for Exceptional Human Longevity. *PLoS Genet.* **11**, 1–23 (2015).
132. Deelen, J. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, (2019).
133. Medici, M. *et al.* Identification of Novel Genetic Loci Associated with Thyroid Peroxidase Antibodies and Clinical Thyroid Disease. *PLoS Genet.* **10**, e1004123 (2014).
134. Jazwinski, S. M. & Kim, S. Examination of the dimensions of biological age. *Frontiers in Genetics* **10**, (2019).
135. Thompson, D. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood is associated with genomic instability in other tissues and susceptibility to non-haematological cancers. *bioRxiv* 514026 (2019). doi:10.1101/514026
136. Sebastiani, P. *et al.* Biomarker signatures of aging. *Aging Cell* **16**, 329–338 (2017).
137. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 14617 (2017).
138. Lu, A. T. *et al.* GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nat. Commun.* **9**, (2018).
139. Suzuki, M. M. & Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
140. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).
141. de Lange, T. How Telomeres Solve the End-Protection Problem. *Science (80-.).* **326**, 948–952 (2009).
142. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat. Med.* **12**, 1133–1138 (2006).
143. Blasco, M. A. The epigenetic regulation of mammalian telomeres. *Nat Rev Genet* **8**, 299–309 (2007).
144. Fagagna, F. d'Adda di *et al.* A DNA damage checkpoint response in telomere-initiated senescence. *Nature* **426**, 194–198 (2003).

145. Takai, H., Smogorzewska, A. & de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr. Biol.* **13**, 1549–56 (2003).
146. Allsopp, R. C. *et al.* Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl. Acad. Sci.* **89**, 10114–10118 (1992).
147. O'Sullivan, R. J. & Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat. Rev. Mol. Cell Biol.* **11**, 171 (2010).
148. Wang, C. & Meier, U. T. Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J.* **23**, 1857–1867 (2004).
149. Bischoff, C. *et al.* The Heritability of Telomere Length Among the Elderly and Oldest-Old. *Twin Res Hum Genet* **8**, 433–439 (2005).
150. Vasa-nicotera, M. *et al.* Mapping of a Major Locus that Determines Telomere Length in Humans. *Am. J. Hum. Genet.* **76**, 147–151 (2005).
151. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.* **45**, 422 (2013).
152. Codd, V. *et al.* Common variants near TERC are associated with mean telomere length. *Nat. Genet.* **42**, 197–199 (2010).
153. Mangino, M. *et al.* DCAF4, a novel gene associated with leucocyte telomere length. *J. Med. Genet.* **52**, 157–162 (2015).
154. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.* **22**, 5056–5064 (2013).
155. Ding, H. *et al.* Regulation of Murine Telomere Length by Rtel : An Essential Gene Encoding a Helicase-like Protein. *Cell* **117**, 873–886 (2004).
156. Pooley, K. A. *et al.* Telomere Length in Prospective and Retrospective Cancer Case-Control Studies. **78**, 3170–3177 (2010).
157. Gu, J. *et al.* A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte telomere length and as a marker of susceptibility for bladder cancer. *Cancer Prev. Res.* **4**, 514–521 (2011).
158. Haycock, P. C. *et al.* Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study. *J. Am. Med. Assoc. Oncol.* **3**, 636–651 (2017).
159. Shay, J. W. & Wright, W. E. Telomeres and telomerase: three decades of progress.

Nat. Rev. Genet. **20**, 299–309 (2019).

160. Ryder, H. *et al.* Obesity , cigarette smoking , and telomere length in women. *Lancet* **366**, 662–664 (2005).
161. Müezzinler, A. *et al.* Body mass index and leukocyte telomere length dynamics among older adults : Results from the ESTHER cohort. *EXG* **74**, 1–8 (2016).
162. Wulaningsih, W., Kuh, D., Wong, A. & Hardy, R. Adiposity, Telomere Length, and Telomere Attrition in Midlife: the 1946 British Birth Cohort. *J Gerontol A Biol Sci Med Sci* **00**, 1–7 (2017).
163. Müezzinler, A. *et al.* Smoking habits and leukocyte telomere length dynamics among older adults : Results from the ESTHER cohort. *Exp Gerontol* **70**, 18–25 (2015).
164. Weischer, M., Bojesen, S. E. & Nordestgaard, B. G. Telomere Shortening Unrelated to Smoking , Body Weight , Physical Activity , and Alcohol Intake : 4 , 576 General Population Individuals with Repeat Measurements 10 Years Apart. *PLoS Genet* **10**, 1–11 (2014).
165. Angela R. Starkweather, PhD, ACNP-BC, CNRN, Areej A. Alhaeeri, BS, Alison Montpetit, PhD, RN, Jenni Brumelle, PhD, Kristin Filler, RN, BS, Marty Montpetit, PhD, Lathika Mohanraj, PhD, Debra E. Lyon, PhD, RN, FNP-BC, FNAP, FAAN, and C. K. J.-C. An Integrative Review of Factors Associated with Telomere Length and Implications for Biobehavioral Research. *Nurs Res* **100**, 130–134 (2014).
166. Lynn F. Cherkas *et al.* The Association Between Physical Activity in Leisure Time and Leukocyte Telomere Length. *J. Am. Med. Assoc.* **168**, 154–158 (2008).
167. Mundstock, E. *et al.* Effect of Obesity on Telomere Length : Systematic Review and. *Obesity* **23**, 2165–2174 (2015).
168. Adler, N. *et al.* NIH Public Access. *Brain Behav Immun* **27**, 15–21 (2014).
169. Kajantie, E. *et al.* No association between body size at birth and leucocyte telomere length in adult life — evidence from three cohort studies. *Int. J Epidemiol* **41**, 1400–1408 (2012).
170. Theall, K. P., Shirtcliff, E. A., Dismukes, A. R., Wallace, M. & Drury, S. S. Association Between Neighborhood Violence and Biological Stress in Children. *J. Am. Med. Assoc. Pediatr.* **171**, 53–60 (2017).
171. Njajou, O. T. *et al.* Telomere length is paternally inherited and is associated with parental lifespan. **104**, 12135–12139 (2007).

172. Burtner, C. R. & Kennedy, B. K. Progeria syndromes and ageing: what is the connection? *Nat. Rev. Mol. Cell Biol.* **11**, 567–578 (2010).
173. Wong, J. M. Y. & Collins, K. Telomere maintenance and disease. *Lancet* **362**, 983–988 (2003).
174. Armanios, M. & Blackburn, E. H. The telomere syndromes. *Nat. Rev. Genet.* **13**, 693–704 (2012).
175. Howlett, N. G. Biallelic Inactivation of BRCA2 in Fanconi Anemia. *Science (80-.).* **297**, 606–609 (2002).
176. Blackburn, E. H., Epel, E. S. & Lin, J. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science (80-.).* **350**, 1193–1198 (2015).
177. Said, M. A., Eppinga, R. N., Hagemeijer, Y., Verweij, N. & van der Harst, P. Telomere Length and Risk of Cardiovascular Disease and Cancer. *J. Am. Coll. Cardiol.* **70**, 506–507 (2017).
178. Fyhrquist, F., Saijonmaa, O. & Strandberg, T. The roles of senescence and telomere shortening in cardiovascular disease. *Nat .Rev. Cardiol.* **10**, 274–283 (2013).
179. Haycock, P. C. *et al.* Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. *BMJ* **349**, (2014).
180. Zhan, Y. *et al.* Telomere Length Shortening and Alzheimer Disease—A Mendelian Randomization Study. *J. Am. Med. Assoc.* **72**, 1202–1203 (2015).
181. Honig, L. S., Kang, M. S., Schupf, N., Lee, J. H. & Mayeux, R. Association of Shorter Leukocyte Telomere Repeat Length With Dementia and Mortality. *J. Am. Med. Assoc. Neurol.* **69**, 1332 (2012).
182. D'Mello, M. J. J. *et al.* Association between shortened leukocyte telomere length and cardiometabolic outcomes: systematic review and meta-analysis. *Circ. Cardiovasc. Genet.* **8**, 82–90 (2015).
183. Forero, D. A. *et al.* Meta-analysis of Telomere Length in Alzheimer's Disease. *J. Gerontol. A. Biol. Sci. Med. Sci.* **71**, 1069–73 (2016).
184. Willeit, P., Willeit, J., Kloss-Brandstatter, Kronenberg, F. & Kiechl, S. Fifteen-Year Follow-up of Association Between Telomere Length and Incident Cancer and Cancer Mortality. *J. Am. Med. Assoc.* **306**, 42–44 (2011).
185. Willeit, P. *et al.* Telomere length and risk of incident cancer and cancer mortality. *J.*

Am. Med. Assoc. **304**, 69–75 (2010).

186. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
187. Graham, M. K. & Meeker, A. Telomeres and telomerase in prostate cancer development and therapy. *Nat. Rev. Urol.* **14**, 607–619 (2017).
188. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Publ. Gr.* **12**, 363–375 (2011).
189. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
190. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
191. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
192. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
193. Forsberg, L. a *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
194. Loh, P. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
195. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
196. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–8 (2016).
197. Acuna-Hidalgo, R. *et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).
198. Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E. & Nickerson, D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40**, 1199–1203 (2008).
199. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).

200. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease-clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
201. Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *Am. J. Hum. Genet.* **98**, 571–578 (2016).
202. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 1–7 (2016).
203. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
204. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
205. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
206. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
207. Dumanski, J. P. *et al.* Mosaic Loss of Chromosome y in Blood Is Associated with Alzheimer Disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).
208. Zhang, C. *et al.* Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Hum. Mol. Genet.* **24**, 5356–5366 (2015).
209. Iles, M. M. *et al.* The Effect on Melanoma Risk of Genes Previously Associated With Telomere Length. *JNCI J. Natl. Cancer Inst.* **106**, (2014).
210. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci. U.S.A* **107**, 9293–8 (2010).
211. Delgado, D. A. *et al.* Genome-wide association study of telomere length among South Asians identifies a second RTEL1 association signal. *J. Med. Genet.* **55**, 64–71 (2018).
212. Langenberg, C. *et al.* Design and cohort description of the InterAct Project: An examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* **54**, 2272–2282 (2011).
213. Langenberg, C. *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLoS Med.* **11**, (2014).
214. Danesh, J. *et al.* EPIC-Heart: The cardiovascular component of a prospective study of nutritional, lifestyle and biological factors in 520,000 middle-aged participants from

10 European countries. *Eur. J. Epidemiol.* **22**, 129–141 (2007).

215. Kristiansson, K. *et al.* Genome-Wide Screen for Metabolic Syndrome Susceptibility Loci Reveals Strong Lipid Gene Contribution But No Evidence for Common Genetic Basis for Clustering of Metabolic Syndrome Traits. *Circ. Cardiovasc. Genet.* **5**, 242–249 (2012).
216. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121–140 (2008).
217. Ikram, M. A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* **32**, 807–850 (2017).
218. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
219. Storey, J. D. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **64**, 479–498 (2002).
220. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
221. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 1–22 (2013).
222. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
223. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
224. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
225. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
226. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
227. Wang, J., Dayem Ullah, A. Z. & Chelala, C. IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome. *Nucleic Acids Res.* **46**, e47–e47 (2018).
228. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825

(2018).

229. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
230. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
231. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
232. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
233. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, (2015).
234. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
235. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109-14 (2012).
236. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691-7 (2011).
237. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
238. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723-9 (2017).
239. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
240. Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat. Commun.* **10**, (2019).
241. Krenciute, G. *et al.* Nuclear BAG6-UBL4A-GET4 Complex Mediates DNA Damage Signaling and Cell Death. *J. Biol. Chem.* **288**, 20547–20557 (2013).
242. Kim, Y. D. *et al.* Metformin Inhibits Hepatic Gluconeogenesis Through AMP-Activated Protein Kinase-Dependent Regulation of the Orphan Nuclear Receptor SHP. *Diabetes* **57**, 306–314 (2008).

243. Irwin, C. R., Hitt, M. M. & Evans, D. H. Targeting Nucleotide Biosynthesis: A Strategy for Improving the Oncolytic Potential of DNA Viruses. *Front. Oncol.* **7**, 229 (2017).
244. Reichard, P. Interactions between deoxyribonucleotide and DNA synthesis. *Annu. Rev. Biochem.* **57**, 349–374 (1988).
245. Pedroza-García, J. A. *et al.* Role of pyrimidine salvage pathway in the maintenance of organellar and nuclear genome integrity. *Plant J.* **97**, 430–446 (2019).
246. Echols, H. & Goodman, M. F. Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* **60**, 477–511 (1991).
247. Bebenek, K., Roberts, J. D. & Kunkel, T. A. The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J. Biol. Chem.* **267**, 3589–3596 (1992).
248. Feng, I. J. & Radivoyevitch, T. SNP-SNP Interactions between dNTP Supply Enzymes and Mismatch DNA Repair in Breast Cancer. in *2009 Ohio Collaborative Conference on Bioinformatics* 123–128 (IEEE, 2009). doi:10.1109/OCCBIO.2009.25
249. Austin, W. R. *et al.* Nucleoside salvage pathway kinases regulate hematopoiesis by linking nucleotide metabolism with replication stress. *J. Exp. Med.* **209**, 2215 LP – 2228 (2012).
250. Franzolin, E. *et al.* The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14272–14277 (2013).
251. Jobert, L. *et al.* The Human Base Excision Repair Enzyme SMUG1 Directly Interacts with DKC1 and Contributes to RNA Quality Control. *Mol. Cell* **49**, 339–345 (2013).
252. de Lange, T. Shelterin-Mediated Telomere Protection. *Annu. Rev. Genet.* **52**, 223–247 (2018).
253. Deng, Z. *et al.* Inherited mutations in the helicase RTEL1 cause telomere dysfunction and Hoyeraal-Hreidarsson syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3408-16 (2013).
254. Giraud-Panis, M.-J., Teixeira, M. T., Géli, V. & Gilson, E. CST Meets Shelterin to Keep Telomeres in Check. *Mol. Cell* **39**, 665–676 (2010).
255. Kim, M. K. *et al.* Regulation of telomeric repeat binding factor 1 binding to telomeres by casein kinase 2-mediated phosphorylation. *J. Biol. Chem.* **283**, 14144–14152 (2008).
256. Lee, S. S., Bohrson, C., Pike, A. M., Wheelan, S. J. & Greider, C. W. ATM Kinase Is

Required for Telomere Elongation in Mouse and Human Cells. *Cell Rep.* **13**, 1623–1632 (2015).

257. Tong, A. S. *et al.* ATM and ATR Signaling Regulate the Recruitment of Human Telomerase to Telomeres. *Cell Rep.* **13**, 1633–1646 (2015).
258. Beneke, S. *et al.* Rapid regulation of telomere length is mediated by poly(ADP-ribose) polymerase-1. *Nucleic Acids Res.* **36**, 6309–6317 (2008).
259. Gomez, M. *et al.* PARP1 Is a TRF2-associated poly(ADP-ribose)polymerase and protects eroded telomeres. *Mol. Biol. Cell* **17**, 1686–96 (2006).
260. Denchi, E. L. & de Lange, T. Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature* **448**, 1068–1071 (2007).
261. Karlseder, J., Broccoli, D., Dai, Y., Hardy, S. & de Lange, T. p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. *Science* **283**, 1321–1325 (1999).
262. van Steensel, B., Smogorzewska, A. & de Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
263. Arnoult, N. & Karlseder, J. Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol* **22**, 859–866 (2015).
264. Collins, K. & Mitchell, J. R. Telomerase in the human organism. *Oncogene* **21**, 564–579 (2002).
265. Blackburn, E. H. & Collins, K. Telomerase: An RNP Enzyme Synthesizes DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003558–a003558 (2011).
266. Stanley, S. E. *et al.* Loss-of-function mutations in the RNA biogenesis factor NAF1 predispose to pulmonary fibrosis-emphysema. *Sci. Transl. Med.* **8**, 351ra107 (2016).
267. Egan, E. D. & Collins, K. Biogenesis of telomerase ribonucleoproteins. *RNA* **18**, 1747–1759 (2012).
268. Nguyen, D. *et al.* A Polyadenylation-Dependent 3' End Maturation Pathway Is Required for the Synthesis of the Human Telomerase RNA. *Cell Rep.* **13**, 2244–57 (2015).
269. Moon, D. H. *et al.* Poly(A)-specific ribonuclease (PARN) mediates 3'-end maturation of the telomerase RNA component. *Nat. Genet.* **47**, 1482–1488 (2015).
270. Boyraz, B. *et al.* Posttranscriptional manipulation of TERC reverses molecular hallmarks of telomere disease. *J. Clin. Invest.* **126**, 3377–3382 (2016).
271. Deng, T. *et al.* TOE1 acts as a 3' exonuclease for telomerase RNA and regulates

telomere maintenance. *Nucleic Acids Res.* **47**, 391–405 (2019).

272. Schilders, G., Raijmakers, R., Raats, J. M. H. & Pruijn, G. J. M. MPP6 is an exosome-associated RNA-binding protein involved in 5.8S rRNA maturation. *Nucleic Acids Res.* **33**, 6795–6804 (2005).
273. Arnér, E. S. & Eriksson, S. Mammalian deoxyribonucleoside kinases. *Pharmacol. Ther.* **67**, 155–86 (1995).
274. Mutahir, Z. *et al.* Thymidine kinase 1 regulatory fine-tuning through tetramer formation. *FEBS J.* **280**, 1531–1541 (2013).
275. Sabini, E., Hazra, S., Ort, S., Konrad, M. & Lavie, A. Structural basis for substrate promiscuity of dCK. *J. Mol. Biol.* **378**, 607–21 (2008).
276. Irwin, C. R., Hitt, M. M. & Evans, D. H. Targeting Nucleotide Biosynthesis: A Strategy for Improving the Oncolytic Potential of DNA Viruses. *Front. Oncol.* **7**, 229 (2017).
277. Carreras, C. W. & Santi, D. V. The Catalytic Mechanism and Structure of Thymidylate Synthase. *Annu. Rev. Biochem.* **64**, 721–762 (1995).
278. Anderson, D. D., Quintero, C. M. & Stover, P. J. Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. *Proc. Natl. Acad. Sci.* **108**, 15163 LP – 15168 (2011).
279. Bester, A. C. *et al.* Nucleotide Deficiency Promotes Genomic Instability in Early Stages of Cancer Development. *Cell* **145**, 435–446 (2011).
280. Chabes, A. *et al.* Survival of DNA Damage in Yeast Directly Depends on Increased dNTP Levels Allowed by Relaxed Feedback Inhibition of Ribonucleotide Reductase. *Cell* **112**, 391–401 (2003).
281. Davidson, M. B. *et al.* Endogenous DNA replication stress results in expansion of dNTP pools and a mutator phenotype. *EMBO J.* **31**, 895 LP – 907 (2012).
282. Blasco, M. A. Telomeres and human disease: ageing, cancer and beyond. *Nat. Rev. Genet.* **6**, 611–622 (2005).
283. Holohan, B., Wright, W. E. & Shay, J. W. Telomeropathies: An emerging spectrum disorder. *J. Cell Biol.* **205**, 289–299 (2014).
284. Sarek, G., Marzec, P., Margalef, P. & Boulton, S. J. Molecular basis of telomere dysfunction in human genetic diseases. *Nat. Struct. Mol. Biol* **22**, 867–874 (2015).
285. Brouilette, S. W. *et al.* Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control

study. *Lancet* **369**, 107–114 (2007).

286. Benetos, A. *et al.* Short Telomeres Are Associated With Increased Carotid Atherosclerosis in Hypertensive Subjects. *Hypertension* **43**, 182–185 (2004).
287. Brouilette, S., Singh, R. K., Thompson, J. R., Goodall, A. H. & Samani, N. J. White Cell Telomere Length and Risk of Premature Myocardial Infarction. *Arterioscler. Thromb. Vasc. Biol.* **23**, 842–846 (2003).
288. Fitzpatrick, A. L. *et al.* Leukocyte Telomere Length and Cardiovascular Disease in the Cardiovascular Health Study. *Am. J. Epidemiol.* **165**, 14–21 (2006).
289. Wentzensen, I. M., Mirabello, L., Pfeiffer, R. M. & Savage, S. A. The Association of Telomere Length and Cancer: a Meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1238–1250 (2011).
290. Zhu, X. *et al.* The association between telomere length and cancer risk in population studies. *Sci. Rep.* **6**, 22243 (2016).
291. Zhan, Y. *et al.* Exploring the Causal Pathway From Telomere Length to Coronary Heart DiseaseNovelty and Significance. *Circ. Res.* **121**, 214–219 (2017).
292. PRENTICE, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11 (1986).
293. Morris, A., Voight, B., Teslovich, T., Ferreira, T. & Segre, A. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. **44**, (2012).
294. White, J. *et al.* Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA Cardiol.* **366**, 1108–1118 (2016).
295. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
296. Prokopenko, I. *et al.* A Central Role for GRB10 in Regulation of Islet Function in Man. *PLoS Genet.* **10**, 1–13 (2014).
297. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–83 (2013).
298. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **00**, 1–9 (2018).

299. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2018).
300. Zheng, J. *et al.* LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
301. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
302. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
303. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
304. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
305. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *Int. J. Epidemiol.* **32**, 1–22 (2003).
306. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
307. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
308. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
309. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. (2018).
310. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

311. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

312. Staley, J. R. *et al.* PhenoScanner: A database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

313. Sanchez-Espiridion, B. *et al.* Telomere Length in Peripheral Blood Leukocytes and Lung Cancer Risk: A Large Case–Control Study in Caucasians. *Cancer Res.* **74**, 2476 LP – 2486 (2014).

314. Stone, R. C. *et al.* Telomere Length and the Cancer–Atherosclerosis Trade-Off. *PLOS Genet.* **12**, e1006144 (2016).

315. Savage, S. A., Gadalla, S. M. & Chanock, S. J. The Long and Short of Telomeres and Cancer Association Studies. *JNCI J. Natl. Cancer Inst.* **105**, 448–449 (2013).

316. McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406 (2008).

317. Speedy, H. E. *et al.* Germ line mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia. *Blood* **128**, 2319–2326 (2016).

318. Rode, L., Nordestgaard, B. G. & Bojesen, S. E. Long telomeres and cancer risk among 95 568 individuals from the general population. *Int. J. Epidemiol.* **45**, 1634–1643 (2016).

319. Landi, M. T. *et al.* A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679–691 (2009).

320. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **18**, 175–186 (2017).

321. Shi, J. *et al.* Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nat. Genet.* **46**, 482–486 (2014).

322. Weller, M. *et al.* Glioma. *Nat. Rev. Dis. Prim.* **1**, 15017 (2015).

323. Walsh, K. M. *et al.* Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk. *Oncotarget* **6**, 42468–77 (2015).

324. Holohan, B. *et al.* Decreasing initial telomere length in humans intergenerationally understates age-associated telomere shortening. *Aging Cell* **14**, 669–677 (2015).

325. Chen, W. *et al.* Longitudinal versus cross-sectional evaluations of leukocyte telomere

length dynamics: age-dependent telomere shortening is the rule. *Journals Gerontol. Ser. A Biomed. Sci. Med. Sci.* **66**, 312–319 (2011).

326. DF, S., JA, B., SC, M. & al, et. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA* **283**, 2008–2012 (2000).
327. Cawthon, R. M. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.* **37**, e21–e21 (2009).
328. De Lucia Rolfe, E. *et al.* Association between birth weight and visceral fat in adults. *Am. J. Clin. Nutr.* **92**, 347–352 (2010).
329. Lindsay, T. *et al.* Descriptive epidemiology of physical activity energy expenditure in UK adults. The Fenland Study. *medRxiv* 19003442 (2019). doi:10.1101/19003442
330. Godino, J. G. *et al.* Effect of communicating genetic and phenotypic risk for type 2 diabetes in combination with lifestyle advice on objectively measured physical activity: protocol of a randomised controlled trial. *BMC Public Health* **12**, 444 (2012).
331. Cawthon, R. M. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.* **37**, 1–7 (2009).
332. Huzen, J. *et al.* Telomere length loss due to smoking and metabolic traits. *J. Intern. Med.* **275**, 155–163 (2014).
333. Dalgård, C. *et al.* Leukocyte telomere length dynamics in women and men: menopause vs age effects. *Int. J. Epidemiol.* **44**, 1688–1695 (2015).
334. Nordfjäll, K. *et al.* The individual blood cell telomere attrition rate is telomere length dependent. *PLoS Genet.* **5**, e1000375 (2009).
335. Verhulst, S., Aviv, A., Benetos, A., Berenson, G. S. & Kark, J. D. Do leukocyte telomere length dynamics depend on baseline telomere length? An analysis that corrects for 'regression to the mean'. *Eur. J. Epidemiol.* **28**, 859–866 (2013).
336. Farzaneh-Far, R. *et al.* Telomere length trajectory and its determinants in persons with coronary artery disease: longitudinal findings from the heart and soul study. *PLoS One* **5**, e8612 (2010).
337. Bendix, L. *et al.* Longitudinal changes in leukocyte telomere length and mortality in humans. *Journals Gerontol. Ser. A Biomed. Sci. Med. Sci.* **69**, 231–239 (2013).
338. Kark, J. D., Goldberger, N., Kimura, M., Sinnreich, R. & Aviv, A. Energy intake and leukocyte telomere length in young adults. *Am. J. Clin. Nutr.* **95**, 479–487 (2012).
339. Farzaneh-Far, R. *et al.* Association of marine omega-3 fatty acid levels with telomeric

aging in patients with coronary heart disease. *JAMA* **303**, 250–257 (2010).

340. García-Calzón, S. *et al.* Dietary inflammatory index and telomere length in subjects with a high cardiovascular disease risk from the PREDIMED-NAVARRA study: cross-sectional and longitudinal analyses over 5 y. *Am. J. Clin. Nutr.* **102**, 897–904 (2015).
341. Eriksson, J. G. *et al.* Higher serum phenylalanine concentration is associated with more rapid telomere shortening in men. *Am. J. Clin. Nutr.* **105**, 144–150 (2016).
342. Soares-Miranda, L. *et al.* Physical Activity, Physical Fitness and Leukocyte Telomere Length: the Cardiovascular Health Study. *Med. Sci. Sports Exerc.* **47**, 2525 (2015).
343. Van Ockenburg, S. L., de Jonge, P., Van der Harst, P., Ormel, J. & Rosmalen, J. G. M. Does neuroticism make you old? Prospective associations between neuroticism and leukocyte telomere length. *Psychol. Med.* **44**, 723–729 (2014).
344. Van Ockenburg, S. L. *et al.* Stressful life events and leukocyte telomere attrition in adulthood: a prospective population-based cohort study. *Psychol. Med.* **45**, 2975–2984 (2015).
345. Dowd, J. B. *et al.* Persistent herpesvirus infections and telomere attrition over 3 years in the Whitehall II cohort. *J. Infect. Dis.* **216**, 565–572 (2017).
346. Ferreira, M. S. V. *et al.* Evidence for a pre-existing telomere deficit in non-clonal hematopoietic stem cells in patients with acute myeloid leukemia. *Ann. Hematol.* **96**, 1457–1461 (2017).
347. Townsley, D. M. *et al.* Danazol treatment for telomere diseases. *N. Engl. J. Med.* **374**, 1922–1931 (2016).
348. Ping, F. *et al.* Deoxyribonucleic acid telomere length shortening can predict the incidence of non-alcoholic fatty liver disease in patients with type 2 diabetes mellitus. *J. Diabetes Investig.* **8**, 174–180 (2017).
349. Masi, S. *et al.* Rate of telomere shortening and cardiovascular damage: a longitudinal study in the 1946 British Birth Cohort. *Eur. Heart J.* **35**, 3296–3303 (2014).
350. Epel, E. S. *et al.* The rate of leukocyte telomere shortening predicts mortality from cardiovascular disease in elderly men. *Aging (Albany NY)* **1**, 81 (2009).
351. Wang, L., Xiao, H., Zhang, X., Wang, C. & Huang, H. The role of telomeres and telomerase in hematologic malignancies and hematopoietic stem cell transplantation. *J. Hematol. Oncol.* **7**, 61 (2014).
352. Barnett, A. G., van der Pols, J. C. & Dobson, A. J. Regression to the mean: what it is

and how to deal with it. *Int. J. Epidemiol.* **34**, 215–220 (2004).

353. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).
354. Lleo, A. *et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *J. Autoimmun.* **41**, 87–91 (2013).
355. Persani, L. *et al.* Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis. *J. Autoimmun.* **38**, J193–J196 (2012).
356. Haitjema, S. *et al.* Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid Endarterectomy. *Circ. Cardiovasc. Genet.* **10**, (2017).
357. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
358. Zhou, W. *et al.* Reply to 'Mosaic loss of chromosome Y in leukocytes matters'. *Nat. Genet.* **51**, 7–9 (2019).
359. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).
360. Bonnefond, A. *et al.* Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
361. Zimmet, P., Alberti, K. G. M. M. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**, 782–787 (2001).
362. Barbieri, M., Bonafè, M., Franceschi, C. & Paolisso, G. Insulin/IGF-I-signaling pathway: an evolutionarily conserved mechanism of longevity from yeast to humans. *Am. J. Physiol. Metab.* **285**, E1064–E1071 (2003).
363. Tatar, M., Bartke, A. & Antebi, A. The Endocrine Regulation of Aging by Insulin-like Signals. *Science (80-.).* **299**, 1346–1351 (2003).
364. Abbasi, A. *et al.* Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* **345**, e5900 (2012).
365. Kengne, A. P. *et al.* Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): A validation of existing models. *Lancet Diabetes Endocrinol.* **2**, 19–29 (2014).
366. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224

(2018).

367. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).

368. Floegel, A. *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62**, 639–648 (2013).

369. Toledo, E. *et al.* Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care* **39**, 833–846 (2016).

370. Wang, T. J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **17**, 448–453 (2011).

371. Xu, F. *et al.* Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. *J. Clin. Endocrinol. Metab.* **98**, E1060-5 (2013).

372. Lotta, L. A. *et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLoS Med.* **13**, 1–22 (2016).

373. Eastwood, S. V. *et al.* Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS One* **11**, (2016).

374. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

375. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

376. Beulens, J. W. J. *et al.* Alcohol consumption and risk of type 2 diabetes in European men and women: influence of beverage type and body size The EPIC-InterAct study. *J. Intern. Med.* **272**, 358–370 (2012).

377. Sacerdote, C. *et al.* Lower educational level is a predictor of incident type 2 diabetes in European countries: the EPIC-InterAct study. *Int. J. Epidemiol.* **41**, 1162–1173 (2012).

378. The InterAct Consortium. Long-Term Risk of Incident Type 2 Diabetes and Measures of Overall and Regional Obesity: The EPIC-InterAct Case-Cohort Study. *PLOS Med.* **9**, e1001230 (2012).

379. The InterAct Consortium. Mediterranean Diet and Type 2 Diabetes Risk in the

European Prospective Investigation Into Cancer and Nutrition (EPIC) Study. *Diabetes Care* **34**, 1913 LP – 1918 (2011).

380. Ekelund, U. *et al.* Physical activity reduces the risk of incident type 2 diabetes in general and in abdominally lean and obese men and women: the EPIC-InterAct Study. *Diabetologia* **55**, 1944–1952 (2012).
381. Spijkerman, A. M. W. *et al.* Smoking and long-term risk of type 2 diabetes: The EPIC-InterAct study in European populations. *Diabetes Care* **37**, 3164–3171 (2014).
382. Mori, H. *et al.* Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8242–8247 (2002).
383. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
384. Taub, M. A. *et al.* Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed. *bioRxiv* 749010 (2019). doi:10.1101/749010
385. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
386. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, (2018).
387. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
388. Langenberg, C. & Lotta, L. A. Genomic insights into the causes of type 2 diabetes. *Lancet* **391**, 2463–2474 (2018).
389. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
390. Ding, Z., Mangino, M., Aviv, A., Spector, T. & Durbin, R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42**, 7–10 (2014).
391. De Meyer, T. *et al.* Telomere Length as Cardiovascular Aging Biomarker. *J. Am. Coll. Cardiol.* **72**, 805–813 (2018).
392. Minamino, T. *et al.* Endothelial cell senescence in human atherosclerosis: role of telomere in endothelial dysfunction. *Circulation* **105**, 1541–4 (2002).

393. Daniali, L. *et al.* Telomeres shorten at equivalent rates in somatic tissues of adults. *Nat. Commun.* **4**, 1597 (2013).
394. Willeit, P. *et al.* Cellular aging reflected by leukocyte telomere length predicts advanced atherosclerosis and cardiovascular disease risk. *Arterioscler. Thromb. Vasc. Biol.* **30**, 1649–56 (2010).
395. De Meyer, T. *et al.* Systemic telomere length and preclinical atherosclerosis: the Asklepios Study. *Eur. Heart J.* **30**, 3074–3081 (2009).
396. Fernández-Alvira, J. M. *et al.* Short Telomere Load, Telomere Length, and Subclinical Atherosclerosis. *J. Am. Coll. Cardiol.* **67**, 2467–2476 (2016).
397. Bekaert, S. *et al.* Telomere length and cardiovascular risk factors in a middle-aged population free of overt cardiovascular disease. *Aging Cell* **6**, 639–647 (2007).
398. Benetos, A. *et al.* Tracking and fixed ranking of leukocyte telomere length across the adult life course. *Aging Cell* **12**, 615–621 (2013).
399. Park, J.-I. *et al.* Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature* **460**, 66–72 (2009).
400. Endorf, E. B. *et al.* Telomerase Reverse Transcriptase Deficiency Prevents Neointima Formation Through Chromatin Silencing of E2F1 Target Genes. *Arterioscler. Thromb. Vasc. Biol.* **37**, 301–311 (2017).
401. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science (80-.).* **347**, 81 LP – 83 (2015).
402. Wiktor, A. *et al.* Clinical significance of Y chromosome loss in hematologic disease. *Genes, Chromosom. Cancer* **27**, 11–16 (2000).
403. Crowe, F. L. *et al.* Fruit and vegetable intake and mortality from ischaemic heart disease: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Heart study. *Eur. Heart J.* **32**, 1235–1243 (2011).
404. Verhoeven, J. E. *et al.* Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. *Mol. Psychiatry* **19**, 895–901 (2014).
405. Zhao, S. & Fernald, R. D. Comprehensive Algorithm for Quantitative Real-Time Polymerase Chain Reaction. *J. Comput. Biol.* **12**, 1047–1064 (2005).
406. Ma, Q. *et al.* MAGI3 negatively regulates Wnt/beta-catenin signaling and suppresses malignant phenotypes of glioma cells. *Oncotarget* **6**, 35851–65 (2015).
407. Ma, Q. *et al.* MAGI3 Suppresses Glioma Cell Proliferation via Upregulation of PTEN

Expression. *Biomed. Environ. Sci.* **28**, 502–9 (2015).

408. Dell'Angelica, E. C., Mullins, C. & Bonifacino, J. S. AP-4, a novel protein complex related to clathrin adaptors. *J. Biol. Chem.* **274**, 7278–7285 (1999).
409. Hirst, J., Bright, N. A., Rous, B. & Robinson, M. S. Characterization of a fourth adaptor-related protein complex. *Mol. Biol. Cell* **10**, 2787–2802 (1999).
410. Bauer, P. *et al.* Mutation in the AP4B1 gene cause hereditary spastic paraplegia type 47 (SPG47). *Neurogenetics* **13**, 73–76 (2012).
411. Barber, E. K., Dasgupta, J. D., Schlossman, S. F., Trevillyan, J. M. & Rudd, C. E. The CD4 and CD8 antigens are coupled to a protein-tyrosine kinase (p56lck) that phosphorylates the CD3 complex. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 3277–3281 (1989).
412. Iwashima, M., Irving, B. A., van Oers, N. S., Chan, A. C. & Weiss, A. Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science* **263**, 1136–1139 (1994).
413. Sturm, R. A., Cassady, J. L., Das, G., Romo, A. & Evans, G. A. Chromosomal structure and expression of the human OTF1 locus encoding the Oct-1 protein. *Genomics* **16**, 333–341 (1993).
414. Segil, N., Roberts, S. B. & Heintz, N. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* **254**, 1814–1816 (1991).
415. Roberts, S. B., Segil, N. & Heintz, N. Differential phosphorylation of the transcription factor Oct1 during the cell cycle. *Science* **253**, 1022–1026 (1991).
416. Schild-Poulter, C., Shih, A., Yarymowich, N. C. & Hache, R. J. G. Down-regulation of histone H2B by DNA-dependent protein kinase in response to DNA damage through modulation of octamer transcription factor 1. *Cancer Res.* **63**, 7197–7205 (2003).
417. Wysocka, J. & Herr, W. The herpes simplex virus VP16-induced complex: the makings of a regulatory switch. *Trends Biochem. Sci.* **28**, 294–304 (2003).
418. Lupo, B. & Trusolino, L. Inhibition of poly(ADP-ribosyl)ation in cancer: Old and new paradigms revisited. *Biochim. Biophys. Acta - Rev. Cancer* **1846**, 201–215 (2014).
419. Déjardin, J. & Kingston, R. E. Purification of Proteins Associated with Specific Genomic Loci. *Cell* **136**, 175–186 (2009).
420. Liang, Y. *et al.* Association of ACYP2 and TSPYL6 Genetic Polymorphisms with Risk of Ischemic Stroke in Han Chinese Population. *Mol. Neurobiol.* **54**, 5988–5995 (2017).

421. Liu, M. *et al.* Association between single nucleotide polymorphisms in the TSPYL6 gene and breast cancer susceptibility in the Han Chinese population. *Oncotarget* **7**, 54771–54781 (2016).
422. Boulay, J. L., Dennefeld, C. & Alberga, A. The Drosophila developmental gene snail encodes a protein with nucleic acid binding fingers. *Nature* **330**, 395–398 (1987).
423. Hay, R. T. SUMO: A History of Modification. *Mol. Cell* **18**, 1–12 (2005).
424. Jones, a. M. *et al.* TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. *Gut* **61**, 248–254 (2012).
425. Lührig, S. *et al.* Lrrc34, a novel nucleolar protein, interacts with npm1 and ncl and has an impact on pluripotent stem cells. *Stem Cells Dev.* **23**, 2862–74 (2014).
426. Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* **45**, 613–620 (2013).
427. Chow, A., Hao, Y. & Yang, X. Molecular characterization of human homologs of yeast MOB1. *Int. J. cancer* **126**, 2079–2089 (2010).
428. Lai, Z.-C. *et al.* Control of cell proliferation and apoptosis by mob as tumor suppressor, mats. *Cell* **120**, 675–685 (2005).
429. Kerjan, G. *et al.* Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 6766–6771 (2009).
430. Kiss, T., Fayet-Lebaron, E. & Jády, B. E. Box H/ACA Small Ribonucleoproteins. *Mol. Cell* **37**, 597–606 (2010).
431. Kwak, J. E., Wang, L., Ballantyne, S., Kimble, J. & Wickens, M. Mammalian GLD-2 homologs are poly(A) polymerases. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4407–4412 (2004).
432. Glahder, J. A. & Norrild, B. Involvement of hGLD-2 in cytoplasmic polyadenylation of human p53 mRNA. *APMIS* **119**, 769–775 (2011).
433. Wyman, S. K. *et al.* Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res.* **21**, 1450–1461 (2011).
434. Schmidt, C. K. *et al.* Systematic E2 screening reveals a UBE2D-RNF138-CtIP axis promoting DNA repair. *Nat. Cell Biol.* **17**, 1458–1470 (2015).
435. Lehner, B. *et al.* Analysis of a high-throughput yeast two-hybrid system and its use to

predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics* **83**, 153–167 (2004).

436. Tang, W., Kannan, R., Blanchette, M. & Baumann, P. Telomerase RNA biogenesis involves sequential binding by Sm and Lsm complexes. *Nature* **484**, 260–264 (2012).
437. Baumann, P. Pot1, the Putative Telomere End-Binding Protein in Fission Yeast and Humans. *Science (80-.).* **292**, 1171–1175 (2001).
438. Hockemeyer, D. & Collins, K. Control of telomerase action at human telomeres. *Nat. Struct. Mol. Biol.* **22**, 848–852 (2015).
439. Lange, T. De. Shelterin : the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100–2110 (2005).
440. Shimizu, A. *et al.* A novel giant gene CSMD3 encoding a protein with CUB and sushi multiple domains: a candidate gene for benign adult familial myoclonic epilepsy on human chromosome 8q23.3-q24.1. *Biochem. Biophys. Res. Commun.* **309**, 143–154 (2003).
441. Toomes, C. *et al.* The presence of multiple regions of homozygous deletion at the CSMD1 locus in oral squamous cell carcinoma question the role of CSMD1 in head and neck carcinogenesis. *Genes. Chromosomes Cancer* **37**, 132–140 (2003).
442. Scholnick, S. B. & Richter, T. M. The role of CSMD1 in head and neck carcinogenesis. *Genes, chromosomes & cancer* **38**, 281–283 (2003).
443. Otsuka, M., Mizuno, Y., Yoshida, M., Kagawa, Y. & Ohta, S. Nucleotide sequence of cDNA encoding human cytochrome c oxidase subunit VIc. *Nucleic Acids Res.* **16**, 10916 (1988).
444. Kile, B. T. *et al.* The SOCS box: a tale of destruction and degradation. *Trends Biochem. Sci.* **27**, 235–241 (2002).
445. Chen, L.-Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).
446. Chang, C.-W., Hsu, W.-B., Tsai, J.-J., Tang, C.-J. C. & Tang, T. K. CEP295 interacts with microtubules and is required for centriole elongation. *J. Cell Sci.* **129**, 2501–2513 (2016).
447. Wu, X. *et al.* ATM phosphorylation of Nijmegen breakage syndrome protein is required in a DNA damage response. *Nature* **405**, 477 (2000).
448. Banin, S. *et al.* Enhanced Phosphorylation of p53 by ATM in Response to DNA

Damage. *Science (80-.).* **281**, 1674 LP – 1677 (1998).

449. Fan, J. *et al.* Tetrameric Acetyl-CoA Acetyltransferase 1 Is Important for Tumor Growth. *Mol. Cell* **64**, 859–874 (2016).

450. Fukao, T. *et al.* Molecular cloning and sequence of the complementary DNA encoding human mitochondrial acetoacetyl-coenzyme A thiolase and study of the variant enzymes in cultured fibroblasts from patients with 3-ketothiolase deficiency. *J. Clin. Invest.* **86**, 2086–2092 (1990).

451. Liu, L. *et al.* MCAF1/AM is involved in Sp1-mediated maintenance of cancer-associated telomerase activity. *J. Biol. Chem.* **284**, 5165–5174 (2009).

452. Liu, L. *et al.* MCAF1/AM Is Involved in Sp1-mediated Maintenance of Cancer-associated Telomerase Activity. *J. Biol. Chem.* **284**, 5165–5174 (2009).

453. Lee, J. & Zhou, P. DCAFs, the Missing Link of the CUL4-DDB1 Ubiquitin Ligase. *Mol. Cell* **26**, 775–780 (2007).

454. Gao, J. *et al.* The CUL4-DDB1 ubiquitin ligase complex controls adult and embryonic stem cell differentiation and homeostasis. *Elife* **4**, (2015).

455. Axe, E. L. *et al.* Autophagosome formation from membrane compartments enriched in phosphatidylinositol 3-phosphate and dynamically connected to the endoplasmic reticulum. *J. Cell Biol.* **182**, 685 LP – 701 (2008).

456. Shen, Z., Huang, S., Fang, M. & Wang, X. ENTPD5, an Endoplasmic Reticulum UDPase, Alleviates ER Stress Induced by Protein Overloading in AKT-Activated Cancer Cells. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 217–223 (2011).

457. Fang, M. *et al.* The ER UDPase ENTPD5 Promotes Protein N-Glycosylation, the Warburg Effect, and Proliferation in the PTEN Pathway. *Cell* **143**, 711–724 (2010).

458. Heeringa, S. F. *et al.* COQ6 mutations in human patients produce nephrotic syndrome with sensorineural deafness. *J. Clin. Invest.* **121**, 2013–2024 (2011).

459. Tsang, W. Y. *et al.* CP110 Cooperates with Two Calcium-binding Proteins to Regulate Cytokinesis and Genome Stability. *Mol. Biol. Cell* **17**, 3423–3434 (2006).

460. Hayashi, R., Goto, Y., Ikeda, R., Yokoyama, K. K. & Yoshida, K. CDCA4 is an E2F transcription factor family-induced nuclear factor that regulates E2F-dependent transcriptional activation and cell proliferation. *J. Biol. Chem.* **281**, 35633–35648 (2006).

461. Kranz, T. M. *et al.* The chromosome 15q14 locus for bipolar disorder and

schizophrenia: is C15orf53 a major candidate gene? *J. Psychiatr. Res.* **46**, 1414–1420 (2012).

462. Ebinu, J. O. *et al.* RasGRP links T-cell receptor signaling to Ras. *Blood* **95**, 3199–3203 (2000).
463. Roose, J. P., Mollenauer, M., Gupta, V. A., Stone, J. & Weiss, A. A diacylglycerol-protein kinase C-RasGRP1 pathway directs Ras activation upon antigen receptor stimulation of T cells. *Mol. Cell. Biol.* **25**, 4426–4441 (2005).
464. van der Velden, L. M. *et al.* Heteromeric interactions required for abundance and subcellular localization of human CDC50 proteins and class 1 P4-ATPases. *J. Biol. Chem.* **285**, 40088–40096 (2010).
465. Paulusma, C. C. & Oude Elferink, R. P. J. The type 4 subfamily of P-type ATPases, putative aminophospholipid translocases with a role in human disease. *Biochim. Biophys. Acta* **1741**, 11–24 (2005).
466. Gao, L. *et al.* Identification of Rare Variants in ATP8B4 as a Risk Factor for Systemic Sclerosis by Whole-Exome Sequencing. *Arthritis Rheumatol.* **68**, 191–200 (2016).
467. Hosford, D. *et al.* Candidate Single-Nucleotide Polymorphisms From a Genomewide Association Study of Alzheimer Disease. *JAMA Neurol.* **65**, 45–53 (2008).
468. Palfreyman, M. T. & Jorgensen, E. M. Unc13 Aligns SNAREs and Superprimes Synaptic Vesicles. *Neuron* **95**, 473–475 (2017).
469. McRory, J. E. *et al.* Molecular and functional characterization of a family of rat brain T-type calcium channels. *J. Biol. Chem.* **276**, 3999–4011 (2001).
470. Cribbs, L. L. *et al.* Cloning and characterization of alpha1H from human heart, a member of the T-type Ca2+ channel gene family. *Circ. Res.* **83**, 103–109 (1998).
471. Daniil, G. *et al.* CACNA1H Mutations Are Associated With Different Forms of Primary Aldosteronism. *EBioMedicine* **13**, 225–236 (2016).
472. Vitko, I. *et al.* Functional Characterization and Neuronal Modeling of the Effects of Childhood Absence Epilepsy Variants of CACNA1H, a T-Type Calcium Channel. *J. Neurosci.* **25**, 4844–4855 (2005).
473. Van Steensel, B., Smogorzewska, A. & De Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
474. Tian, Y. *et al.* C. elegans Screen Identifies Autophagy Genes Specific to Multicellular Organisms. *Cell* **141**, 1042–1055 (2010).

475. Smogorzewska, A. *et al.* Control of human telomere length by TRF1 and TRF2. *Mol. Cell. Biol.* **20**, 1659–68 (2000).

476. Inano, S. *et al.* RFWD3-Mediated Ubiquitination Promotes Timely Removal of Both RPA and RAD51 from DNA Damage Sites to Facilitate Homologous Recombination. *Mol. Cell* **66**, 622-634.e8 (2017).

477. Fu, X. *et al.* RFWD3-Mdm2 ubiquitin ligase complex positively regulates p53 stability in response to DNA damage. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4579–4584 (2010).

478. Lehner, B. & Sanderson, C. M. A protein interaction framework for human mRNA degradation. *Genome Res.* **14**, 1315–1323 (2004).

479. Shintani, M., Urano, M., Takakuwa, Y., Kuroda, M. & Kamoshida, S. Immunohistochemical characterization of pyrimidine synthetic enzymes, thymidine kinase-1 and thymidylate synthase, in various types of cancer. *Oncol. Rep.* **23**, 1345–1350 (2010).

480. Tempel, W. *et al.* Nicotinamide riboside kinase structures reveal new pathways to NAD+. *PLoS Biol.* **5**, e263 (2007).

481. Han, Z. G. *et al.* Molecular cloning of six novel Krüppel-like zinc finger genes from hematopoietic cells and identification of a novel transregulatory domain KRNB. *J. Biol. Chem.* **274**, 35741–8 (1999).

482. Kotenko, S. V *et al.* IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex. *Nat. Immunol.* **4**, 69–77 (2003).

483. Prosser, H. M. *et al.* Prokineticin receptor 2 (Prokr2) is essential for the regulation of circadian behavior by the suprachiasmatic nuclei. *Proc. Natl. Acad. Sci.* **104**, 648 LP – 653 (2007).

484. Dodé, C. & Rondard, P. PROK2/PROKR2 Signaling and Kallmann Syndrome. *Front. Endocrinol. (Lausanne).* **4**, 19 (2013).

485. Zhu, L. *et al.* Inhibition of cell proliferation by p107, a relative of the retinoblastoma protein. *Genes Dev.* **7**, 1111–1125 (1993).

486. Ryoo, J. *et al.* The ribonuclease activity of SAMHD1 is required for HIV-1 restriction. *Nat. Med.* **20**, 936–941 (2014).

487. Laguette, N. *et al.* SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* **474**, 654–657 (2011).

488. Margalef, P. *et al.* Stabilization of Reversed Replication Forks by Telomerase Drives

Telomere Catastrophe. *Cell* **172**, 439-453.e14 (2018).

489. Ballew, B. J. *et al.* A recessive founder mutation in regulator of telomere elongation helicase 1, RTEL1, underlies severe immunodeficiency and features of Hoyeraal Hreidarsson syndrome. *PLoS Genet.* **9**, e1003695 (2013).
490. Stuart, B. D. *et al.* Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat. Genet.* **47**, 512 (2015).
491. Zhang, Y. *et al.* Overexpression of SCLIP promotes growth and motility in glioblastoma cells. *Cancer Biol. Ther.* **16**, 97–105 (2015).
492. You, R. *et al.* Apoptosis of dendritic cells induced by decoy receptor 3 (DcR3). **111**, 1480–1489 (2019).
493. Pitti, R. M. *et al.* Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396**, 699–703 (1998).
494. Yang, C.-R. *et al.* Soluble decoy receptor 3 induces angiogenesis by neutralization of TL1A, a cytokine belonging to tumor necrosis factor superfamily and exhibiting angiostatic action. *Cancer Res.* **64**, 1122–1129 (2004).
495. Chevrier, S. & Corcoran, L. M. BTB-ZF transcription factors, a growing family of regulators of early and late B-cell development. *Immunol. Cell Biol.* **92**, 481–8 (2014).
496. Chen, W.-Y. *et al.* Inhibition of the androgen receptor induces a novel tumor promoter, ZBTB46, for prostate cancer metastasis. *Oncogene* **36**, 6213 (2017).
497. Li, J. S. Z. *et al.* TZAP: A telomere-associated protein involved in telomere length control. *Science (80-.).* **355**, 638–641 (2017).
498. Jahn, A. *et al.* ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.* **18**, 929–946 (2017).
499. Adamson, B., Smogorzewska, A., Sigoillot, F. D., King, R. W. & Elledge, S. J. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.* **14**, 318–328 (2012).

Appendix A

Supplementary Notes

Information on study cohorts

The demographic characteristics of all study cohorts, for both discovery and replication phases are shown in Supplementary Table 1. All individuals included in the analysis are of European descent.

ENGAGE
The majority of the studies included have previously been described[151]. In addition to these the following studies were included in this analysis.

GENMETS
GENMETS is a subcohort of the Finnish population-based Health 2000 study, comprising of metabolic syndrome cases and controls. This cohort is described in more detail elsewhere[215].

NESDA
The Netherlands Study of Depression and Anxiety (NESDA) is an ongoing cohort study into the long- term course and consequences of depressive and anxiety disorders. A description of the study rationale, design, and methods is given elsewhere[216]. Briefly, in 2004 to 2007, participants aged 18 to 65 years were recruited from the community (19%), general practice (54%), and secondary mental health care (27%), therefore reflecting various settings and developmental stages of psychopathology to obtain a full and generalizable picture of the course of psychiatric disorders. A total of 2981 participants were included, consisting of persons with a current or past depressive and/or anxiety disorder and healthy control subjects. Exclusion criteria were a clinically overt primary diagnosis of psychotic, obsessive compulsive, bipolar, or severe addiction disorder and not being fluent in Dutch. The research protocol was approved by the ethical committee of participating universities, and all respondents provided written informed consent.

ROTTERDAM
The Rotterdam Study is a population-based cohort study that investigates the occurrence and determinants of diseases in the elderly, which has been ongoing since 1990[217]. As of 2008, detailed phenotypic and genetic data has been collected on ~15,000 subjects aged 45 years or over. For this study the RS-I and RS-III cohorts were used. The Medical Ethics Committee at Erasmus Medical Centre approved the study protocol.

EPIC-InterAct case-cohort study
The EPIC-InterAct study aimed to investigate the independent and interactive effects of genetic and behavioural risk factors on type 2 diabetes risk[212,213]. EPIC-InterAct is a case-cohort study nested within 8 of the 10 countries participating in the EPIC-Europe cohort study. EPIC-InterAct ascertained 12,403 cases of type 2 diabetes from a total cohort of 340,234 participants who provided blood samples at baseline and were followed-up for an average of

7 years (~4 million years of follow-up). Cases were ascertained from multiple data sources including self-report of a physician diagnosis of diabetes, linkage to primary/secondary care records, medication use, hospital admission data and death registration data. We also established a random sub-cohort of 16,154 participants who were representative of participants within each country. By design there is an overlap with the set of incident diabetes cases (n=778). Participant characteristics have been previously reported in detail[212,213]. Observational statistics of LTL, genotyping and imputation are summarised in the Supplementary Table 1 and 2.

EPIC-CVD case-cohort study

EPIC-CVD was designed as a case-cohort study that uses the same random sub-cohort as InterAct, with a focus on incident coronary heart disease and stroke events[214]. The participants included in this analysis are thus incident cases only (7722 coronary heart disease cases and 3451 cerebrovascular disease cases). We also included an additional 752 participants as a random sub-cohort from the two countries not included in EPIC-InterAct (Greece and Norway). Detailed characteristics of the EPIC-CVD participants has been previously reported[403].

Telomere length measurements

Telomere length measurements were performed using an established quantitative PCR technique[327] across 6 laboratories. Laboratory specific information is given below and in Table S1. Details of the techniques used within Helsinki, Leicester and London have been given elsewhere[151].

NESDA: Fasting blood was drawn from participants in the morning between 8:30 and 9:30 am and blood samples were stored in a -80°C freezer afterwards. Leukocyte TL was determined at the laboratory of Telomere Diagnostics, Inc. (Menlo Park, CA, USA), using quantitative polymerase chain reaction (qPCR), adapted from the published original method[327]. Telomere sequence copy number in each patient's sample (T) was compared to a single-copy gene copy number (S), relative to a reference sample. The detailed method is described elsewhere[404].

Rotterdam: Telomere length was measured using a qPCR assay based on the method described elsewhere[327]. with minor modifications. For each sample the telomere and 36B4 assay were run in separate wells but in the same 384 wells PCR plate. Each reaction contained 5ng DNA, 1uM of each of the telomere primers (tel1b-forward: GGTTTGTTTGGGTTTGGGTTTGGGTTTGGGTTTGGGTT, tel2b-reverse: GGCTTGCCTTACCCTTACCCTTACCCTTACCCTTACCCT) or 250nM of the 34B4 primers (36B4u-forward: CAGCAAGTGGGAAGGTGTAATCC, 36B4d-reverse: CCCATTCTATCATCAACGGGTACAA) and 1x Quantifast SYBR green PCR Mastermix (Qiagen). The reactions for both assays were performed in duplicate for each sample in a 7900HT machine (Applied Biosystems). Ct values and PCR efficiencies were calculated per plate using the MINER algorithm[405]. Duplicate Ct values that had a Coefficient of Variance (CV) of more than 1% were excluded from further analysis. Using the average Ct value per sample and the average PCR efficiency per plate the samples were quantified using the formula Q=1/(1+PCR eff)^Ct. The relative telomere length

was calculated by dividing the Q of the telomere assay by the Q of the 34B4 assay. To validate the assay 96 random samples were run twice and the CV of that experiment was 4.5%.

Cambridge: Relative mean LTL was measured using a ViiATM Real-Time quantitative PCR system (ThermoFisher Sicentific, Inc), and expressed as a ratio (T/S) of the relative quantities of the telomeric TTAGGG repeat (T) and the single copy of a housekeeping gene, *Albumin* (S). The denominator determines total genome copies per sample, controlling the technical errors during quantification. The measurement was validated by the Terminal Restriction Fragment (TRF) analysis (the "gold standard" measurement of TL) using separate DNA samples extracted from peripheral blood mononuclear cells in 30 individuals (Pearson's r =0.69). Batch effect was corrected by normalising all the other batches to the fourth batch. Each sample was measured repetitively for three times within one batch, when the same sample was measured in more than one batch, measurement from the last batch was kept for the sample. Samples with coefficients of variation greater than 10% were excluded.

Description of Individual loci associated with LTL

Chr1p13.2. The lead SNP (rs12065882) and three high LD variants are all located within introns of *MAGI3* (*membrane associated guanylate kinase, WW and PDZ domain containing 3*). *MAGI3* has been proposed to act as a tumour suppressor; it regulates cell proliferation in glioma via wnt/β-catenin signalling and interacts with PTEN[406,407]. Both S-PrediXcan and COLOC analyses give evidence to support expression of *AP4B1* (*adaptor related protein complex 4 subunit beta 1*) being influenced by the associated variants. This gene encodes a subunit of a heterotetrametric adapter-like complex 4 that involves in Golgi-associated and lysosomal vesicle biogenesis and membrane trafficking, transporting proteins from the trans-Golgi network to the endosomal-lysosomal system[408,409]. Mutations in this gene are associated with an autosomal recessively inherited disease, spastic paraplegia type 47[410]. There is also evidence of a colocalised eQTL signal for *PTPN22* (*protein tyrosine phosphatase, non-receptor type 22*) in three tissues. PTPN22 interacts with the proto-oncogene CBL, a member of the E3 ubiquitin ligase family that has been implicated in several cancers.

Chr1q24.2. rs35675808 is located downstream of the 3' UTR of *CD247 (CD247 molecule),* which encodes T-cell receptor zeta that constitutes the T-cell receptor-CD3 complex, coupling antigen recognition to several signalling transduction pathways, essential in adaptive immune response[411,412]. Pathways that have been shown to be implicated with this gene include HIV life cycle and translocation of ZAP-70 to immunological synapse (Reactome). Mutations in this gene are associated with autosomal recessive immunodeficiency 25, characterised by T-cells impaired response to alloantigens, tetanus toxoid and mitogens (OMIM #610163). Another gene, the *POU2F1* (*POU class 2 homeobox 1*), located 3kb upstream of this variant, might be biologically relevant. This gene, also known as the *OCT1*, belongs to the first identified members of the POU transcription factor family[413,414]. Members of this family contain the POU domain, a 160-amino acid region necessary for DNA binding to the octameric motif (5'-ATGCAAAT-3') (OMIM #164175). POU2F1, as a transcriptional factor, is involved in cell cycle regulation and transcription of histone H2B and other cellular housekeeping genes[414,415]. It has also been suggested that the expression of histone H2B was downregulated in response to double-stranded DNA breaks via a mechanism that modulates transcriptional regulatory potential of POU2F1 by site-specific phosphorylation[416]. POU2F1 is implicated with various pathways, including the RNA Polymerase III transcription initiation, cytokine signalling in immune system, BRCA1 pathway and glucocorticoid receptor signalling (Reactome). This gene also facilitates human herpes simplex virus (HSV) infection by forming a multiprotein-DNA complex with the virion proteins, activating transcription of the viral immediate early genes[417].

Chr1q42.12. Variants at this locus are focused across the *PARP1* gene, which encodes the first protein member of the poly(ADP-ribosyl)transferases family, also termed as the ADP-ribosyltransferases with diphtheria toxin homology (ARTDs). It plays an essential role in various pathways of DNA repair and chromatin remodelling, including single- and double-strand break repair, nucleotide excision repair, stabilization of replication forks, and modulation of chromatin structure, thereby maintaining genomic integrity and stability[418]. Because the DNA double-strand breaks structurally resemble telomeres, regulators and components of DNA repair machinery have been shown to be implicated in telomere homeostasis[263]. Of note, rs1136410 (r^2=1.0 to the lead) causes a known V762A substitution in PARP1 (poly(ADP-ribose) polymerase 1), which has been shown to reduce PARP1 activity. The

allele that reduces activity is associated with shorter LTL, consistent with previous studies where knockdown of PARP1 leads to telomere shortening. PARP1 was identified as a telomeric double-stranded repeats binding factor in a proteomic study of telomeres using DNA *in situ* hybridization in conjugation with mass spectrometry[419]. PARP1 poly(ADP-ribosyl)ates TRF2, which affects TRF2 binding to the telomere (Gomez et al., 2006). In addition to the coding change there is also eQTL evidence for *PARP1* (S-PrediXcan and COLOC, online methods, Supplementary table 7) in pancreas, with the shorter LTL allele associating with reduced *PARP1* expression. Another SNP, rs907187, is highlighted in the integrated analysis of non-coding variants and is located within the 5' UTR of *PARP1*, which could mediate the effect on gene expression.

Chr2p16.2. rs754017156 is located within intron 3 of *ACYP2* (*acylphosphatase 2*) and also causes an in-frame insertion of two amino acids into *TSPYL6* (*TSPY like 6*). This gene encodes a nuclear protein, the Testis-Specific Y-Encoded-Like Protein 6, that involves in the nucleosome assembly. Biological function of this protein is largely unexplored. Studies have associated genetic polymorphisms of this gene region with increased risk of ischemic stroke[420], and breast cancer in the Han Chinese population[421]. There are no high LD SNPs, but an evidence of an eQTL in testis for *TSPYL6*.

Chr2q34. rs56810761 is located within intron 7 of *UNC80* (*unc-80 homolog, NALCN channel complex subunit, A*) gene. There are no high LD SNPs, but an evidence of an eQTL for *SNAI1P1* (*snail family zinc finger 1 pseudogene 1*) in testis in the co-localisation analysis. *SNAIP1* is a processed pseudogene of *SNAI1*, which encodes the human ortholog of a zinc finger protein of the snail family, first cloned in *Drosophila*, which was demonstrated to be essential in the formation of mesoderm during gastrulation and embryonic development[422].

Chr3q12.3. This locus consists of a 77 SNPs located predominantly across *SENP7* (*SUMO1/sentrin specific peptidase 7*) gene. The lead SNP is located 53bp upstream of *SENP7* within a proximal promoter. It is associated with a DNaseI sensitivity QTL and with *SENP7* expression in one tissue (co-localisation). Lower expression of *SENP7* associates with shorter LTL. Although it has no known role in telomere regulation, the small ubiquitin-like modifier (SUMO) functions as a post-translational modification, regulating various biological events, especially in DNA repair, chromatin organization, transcription, and RNA metabolism[423], which are essential biological events pertinent to telomere homeostasis.

Chr3q13.2. The variants in this region are all located within intron 2 of a predicted mRNA, *RP11-572M11.4* and downstream of a non-coding RNA *RP11-572M11.3* (also named *LINC02044*). There is no supporting evidence to suggest which gene is potentially influenced at this locus.

Chr3q26.2. This locus contains 47 SNPs in high LD ($r^2 < 0.8$) with the lead SNP (rs1093660). The *telomerase RNA component* (*TERC*) is the functional candidate in this locus. One SNP (rs2293607, r^2=0.81 to rs1093660) is located 63bp downstream of the *TERC* sequence, which potentially leads to altered *TERC* expression[424]. However, the lead variant, rs10936600, encodes a L241I substitution within *LRRC34* (*Leucine rich repeat containing 34*), which is predicted to be deleterious (Supplementary Table 6). The CADD score (19.81) places this SNP

just outside of the 1% most deleterious mutations. LRRC34 is a member of the leucine rich repeat containing protein family. Although little is known about its biological function, it has been suggested to be implicated in the maintenance and regulation of pluripotency[425]. Knock down of LRRC34 results in reduced expression of some, but not all, pluripotency genes[425]. As genes encoding the telomerase enzyme share the same expression patterns as those of the pluripotency genes, thereby they are potentially subjected to the LRRC34-mediated transcriptional regulation. Another highly linked variant, rs10936599 (r^2=1.0) is predicted to have a functional effect in the integrated analysis of non-coding variants (Supplementary Table 7). It is located on the edge of the active promoter region of *MYNN*, just inside the coding sequence. An eQTL is observed for *MYNN* in testis (shorter TL associated with higher expression), suggesting that this SNP may alter *MYNN* expression. MYNN protein is a member of the BTB/POZ and zinc finger containing family that is involved in transcriptional regulation. It has also been shown to interact with CUL3, a core component of the E3 Ubiquitin ligase complex, which functions in many cellular processes including DNA repair. LTL variants at this locus have been associated with idiopathic pulmonary fibrosis, of which telomere dysregulation is attributed to the disease aetiology[426]. Despite the obvious involvement of *TERC* in telomere length regulation, little bioinformatic evidence is available to support it to be the only likely-causal gene in this region, i.e. other candidate genes might also explain the locus association, such as *LRRC34* and *MYNN*. However, it is also possible that with *TERC* being a processed non-coding RNA, the relevant information is limited in standard datasets. There are no eQTLs for *TERC* in the GTex dataset, but a study has shown that variants in the regulatory region can affect its expression level, possibly by facilitating the maturation of *TERC* via 3' processing[424].

Chr4q13.3. The lead variant rs13137667 is located within the first intron of *MOB1B* (*MOB kinase activator 1B*). There are 49 variants in high LD, the majority of which are located intronically within *MOB1B* or *DCK* (*deoxycytidine kinase*). No high LD non-synonymous variants or co-localised eQTLs were found at this locus. MOB (Mps one binder) was originally identified as an Mps1 binding protein in yeast, regulating mitotic checkpoint and cytokinesis, and is evolutionarily conserved across all major kingdoms[427]. Human MOB1B homolog activates LATS1/2 (Large tumour suppressor 1/2) through protein-protein interaction in the Hippo signalling pathway, resulting in the inhibition of cell proliferation, apoptosis, and thus tumour suppression[428]. DCK is a key component of the deoxyribonucleoside salvage pathway and phosphorylates deoxycytidine, deoxyguanosine and deoxyadenosine to dCMP, dGMP and dAMP respectively.

Chr4q31.23. There are 65 associated variants clustered towards the 5' end of *DCLK2* (*doublecortin like kinase 2*). There is an eQTL co-localised with *DCLK2* in one tissue (Supplementary table 7). *DCLK2* encodes a protein that contains four independent functional domains: two doublecortin domains at the N-terminus, essential for microtubule binding and regulating microtubule polymerisation, a serine/threonine protein kinase domain at the C-terminus, sharing substantial homology to Ca^{2+}/calmodulin-dependent protein kinase, and a serine/proline-rich domain in between the two termini, which mediates multiple protein-protein interactions. Mouse models with single or double copies of *Dclk2* gene ablated are viable and fertile, however, a simultaneous deletion of *Dcx* gene, encoding another protein member of the doublecortin family, results in spontaneous seizures, hippocampal

disorganisation and poor survival[429], phenotypically mimicking human X-linked lissencephaly (OMIM #613166).

Chr4q32.2. This locus contains 70 closely related (r^2>0.8) SNPs spanning *NAF1* (*nuclear assembly factor 1 ribonucleoprotein*), a gene encoding an RNA-binding protein, required for the synthesis of box H/ACA RNAs and sequential assembly with proteins to form ribonucleoprotein (RNP) complex. The box H/ACA RNPs regulates three fundamental cellular processes: protein synthesis, mRNA splicing via site-specific pseudouridylation of ribosomal RNAs and small nuclear RNAs and telomere maintenance by facilitating the maturation of *TERC* in telomerase[430]. Expression evidence was found for *NAF1* (S-PrediXcan and COLOC) and an antisense transcript *RP11-563E2.2* (COLOC, online methods, Supplementary Table 7). The lead SNP, rs4691895, is a non-synonymous variant in *NAF1* (L368V) along with another high LD variant (rs4691896, r^2=1, I162V). Individually both are predicted to be benign; however, it is unclear what effects they may have in combination.

Chr5p15.33. There are two independently associated SNPs at this locus, neither of which have any high LD variants. Both SNPs are located within intron 2 of *TERT*, but little functional evidence was found to support their involvements in regulating *TERT* levels, which might be due to the transcriptional repression of *TERT* in most somatic tissues.

Chr5q14.1. The lead variant, rs62365174, is located in intron 4 of *TENT2* (*terminal nucleotidyltransferase 2*, previously named *PAPD4 and GLD2*). There are 137 SNPs in high LD (r^2<0.8), which fall across the region of *TENT2* and include upstream, intronic and 3' UTR variants. There is strong evidence that these variants can affect the expression of *TENT2*, with eQTLs co-localised in 9 tissues, exhibiting consistent positive correlations, i.e. reduced expression associates with decreased LTL. TENT2 functions as the cytoplasmic poly(A) RNA polymerase that adds successive AMP monomers to the 3'-end of specific RNAs, forming a poly(A) tail, exhibiting strict substrate specificity, that, different from the canonical nuclear poly(A) RNA polymerase, only functions on cytoplasmic RNAs[431]. Previous studies have suggested its role in the polyadenylation and stability of p53 mRNA[432] and several miRNAs[433].

Chr5q31.2. The associated variant, rs112347796, has no further variants in high LD (r^2>0.8). It is located within intron 1 of *UBE2D2* (*ubiquitin conjugating enzyme E2 D2*), which is involved in the DNA damage repair[434]. There is no evidence to suggest the potential function of this variant.

Chr6p22.2. This locus contains 10 SNPs in high LD (r^2>0.8) with the lead SNP, all located around *CARMIL1* (*capping protein regulator and myosin 1 linker 1*, previously named *LRR16A*). One SNP, rs913455, causes a synonymous change within exon 3 and has scored to have possible regulatory function (Supplementary Table 8), which may be driven in part by its high conservation and location within the coding region. There is no supporting literature evidence to identify which gene(s) may be influenced at this locus.

Chr6p21.33. There are 11 SNPs in high LD (r^2>0.8) with the lead SNP, which are located across the major histocompatibility complex (MHC) class III region. MHC is a highly polymorphic and gene-dense region with complex linkage disequilibrium structure, and thus characterisation of potential causal genes within this region is difficult. A number of genes can potentially

serve as causal gene candidates, including *PRRC2A*, *CSNK2B* and *BAG6*. There is evidence that the expression of both *BAG6* and *CSNK2B* (S-PrediXcan and COLOC, Supplementary Table 7) is affected. The lead variant is located upstream of *PRRC2A*, which was previously known as the *BAT2* (*HLA-B associated transcript 2*) gene, encoding a large protein (2157 amino acids). PRRC2A has been shown to be involved in the pre-mRNA editing, as spliceosome and splicing regulators were found to be able to bind to the PRRC2A in protein-protein interaction assays, including the heterogeneous nuclear RNPs and the cleavage and polyadenylation specific factor 1[435]. As maturation of the telomerase RNA subunit involves a spliceosome-mediated single cleavage reaction[436], PRRC2A may regulate telomere length via involvement in the biogenesis of *TERC*. Of note, another variant, rs805299 (r^2=1), located within intron 1 of *BAG6* (*BCL2 associated athanogene 6*), shows a high probability for promoter activity and is predicted to have regulatory function in the integrated analysis of non-coding variants (Supplementary Table 8). *BAG6* was part of a cluster of genes that encode a multifunctional protein, involved in various pathways, including intracellular protein quality controls by promoting proteasomal degradation of misfolded and mislocalised proteins, and DNA damage-induced apoptosis. Another variant, rs5872 (r^2=1), is located within the 3'UTR of *CSNK2B* (*casein kinase 2 beta*). CSNK2B is a subunit of CSNK2 that is involved in multiple pathways but of note has been shown to interact with TRF1. CSNK2-mediated phosphorylation of TRF1 is required for the binding of TRF1 to telomeres, which has been proposed to be essential for telomere length homeostasis[255].

Chr7q31.33. The associated variants cover the *POT1* (*protection of telomeres 1*) gene, which encodes the most conserved protein component of the shelterin complex among all eukaryotes[437]. It is tethered to the TERF1 and TERF2 homodimers via a TIN2-mediated linkage, and specifically bound to the single-stranded telomeric repeats, protecting it from nucleolytic degradation[438]. Moreover, POT1 controls the sequence precision at the 5' ends, which are identical among nearly all human chromosomes, and regulates telomere length by restricting telomerase binding[439]. Rare nonsense mutations within this gene, which blocked physical interactions of POT1 with telomeric single-stranded repeats and other components of the shelterin protein complex, were identified by whole-exome sequencing in families with strong histories of chronic lymphocytic leukaemia[317]. The integrated analysis of non-coding variants highlights rs2239532 (r^2=0.85), located within the 5'UTR of *GPR37* (*G protein-coupled receptor 37*), as having regulatory function (Supplementary Table 8). Although no direct eQTL evidence is available to support POT1, there is evidence to link the expression of an uncharacterised *POT1-AS* transcript (*RP11-3B12.1*) to LTL via co-localisation in two tissues (Supplementary Table 8).

Chr8p23.2. This region contains 52 SNPs in high LD (r^2<0.8) and is located within 3 introns towards the 3' end of *CSMD1* (*CUB and Sushi multiple domains 1*) gene. *CSMD1* was potentially associated with a rare neurological disease, the benign adult familial myoclonic epilepsy[440]. It may also act as a suppressor of squamous cell carcinomas, yet unequivocal evidence is lacking[441,442]. The gene-knockout mouse was used as a schizophrenia human disease model, exhibiting increased levels of exploratory activity, behavioural despair anxiety-related response, and decreased startle reflex (MGI: 3528558). However, no direct supporting evidence is available to suggest CSMD1 or other genes as causal gene candidates in this region.

Chr8q22.2. Four SNPs are located upstream of *COX6C* (*cytochrome c oxidase subunit 6C*). COX6C is a subunit of complex IV that catalyses the final step of the mitochondrial respiratory chain[443]. No functional data is available to pinpoint causal genes for this locus.

Chr10p15.1. The 6 associated variants (in LD, r^2>0.8) at this locus are clustered within the first intron of *ASB13* (*ankyrin repeat and SOCS box containing 13*), a member of the suppressor of cytokine signalling box protein superfamily. Members of this protein family can also be components of E3 ubiquitin ligase complexes[444]. No causal gene candidates can be prioritised for this locus.

Chr10q24.33. This region contains *STN1* (*STN1, CST complex subunit*, also termed *OBFC1* in humans), a component of the telomere binding CST complex. There is strong evidence that the variants affect *STN1*(*OBFC1*) expressions across multiple tissues (S-PrediXcan and COLOC, Supplementary Table 7). The CST complex regulates telomere maintenance by mediating the access to telomeres for telomerase and DNA polymerase α[445].

Chr11q21. The lead variant, rs117037102, is located within intron 5 of *CEP295* (*centrosomal protein 295*, also termed *KIAA1731*). There is a potentially damaging protein coding variant (rs117405490, r^2=1), which results in a P to A substitution at position 783 of CEP295. CEP295 is a centriole-enriched microtubule-binding protein, highly conserved across species and involved in centriole biogenesis, essential for cell cycle regulation and mitotic progression[446].

Chr11q22.3. The associated variants fall across a ~321kb region which includes several genes, including *ATM* (*ATM serine/threonine kinase*), encoding a protein kinase that phosphorylates many checkpoint-determining and regulatory proteins, such as p53, Chk2 and BRCA1, and thus playing an essential role in cell cycle control and DNA-damage-activated signalling pathways[447]. ATM is responsible for the human genetic disorder ataxia telangiectasia (A-T), manifested with genome instability, cerebellar and thymic degeneration, immunodeficiency, premature ageing, sensitivity to ionizing radiation and predisposition to cancer (OMIM #208900)[448]. There are eQTLs supporting *ATM* and another gene, *ACAT1* (*acetyl-CoA acetyltransferase 1*), within the region. ACAT1 is a mitochondrial protein, expression levels of which have been linked to some cancers[449]. Defects in this gene are associated with 3-ketothiolase deficiency, an inborn error of isoleucine catabolism[450].

Chr12p13.1. The lead variant and 2 in high LD (r^2<0.8) are located upstream of *ATF7IP* (*activating transcription factor 7 interacting protein*), also named *MCAF1*, actively involved in histone modification, chromatin organisation, and Sp1-dependent maintenance of telomerase activity in cancer cells[451]. It was previously shown to regulate expression of both *TERT* and *TERC* and consequently telomerase activity[452].

Chr12q13.13. There are 7 variants in high LD (r^2 <0.8), located within a 3kb region upstream of *SMUG1* (*single-strand-selective monofunctional uracil-DNA glycosylase 1*), a gene involved in base-excision repair. Although there is no bioinformatic evidence to show that these variants affect SMUG1 expression levels, previous functional studies have suggested that SMUG1 might influence telomere length by interacting with the telomerase component Dyskerin (DKC1) with which it controls rRNA processing[251].

Chr14q24.2. The lead variant is a non-synonymous (W22C) variant in *DCAF4* (*DDB1 and CUL4 associated factor 4*). Another variant in high LD (rs3815460, r^2=1) also causes a protein coding change (S345C). Both variants are predicted to be damaging individually. DCAF4 interacts with the Cul4-Ddb1 E3 ubiquitin ligase macromolecular complex, which regulates processes including DNA repair and cellular proliferation[453]. *DDB* (*DNA damage binding protein*) is highly expressed in multipotent hematopoietic progenitors, conditional ablation of which in hematopoietic stem and progenitor cells led to a complete loss of pluripotency and self-renewal of progenitors and stem cells, suggesting its role in cell differentiation, apoptosis and death[454]. An intronic G-to-A variant (rs2535913) has been associated with shorter LTL[153]. A further SNP, rs2286838 (r^2=0.9) causes a coding change in *ZFYVE1* (*zinc finger FYVE-type containing 1*, S408R), which also has a predicted damaging effect. This protein, also known as the *double FYVE-containing protein 1* (*DFCP1*), contains two zinc-binding FYVE domains in tandem, which has been shown to be localised on endoplasmic reticulum and Golgi apparatus via binding to phosphatidylinositol 3-phosphate containing membranes, essential for the regulation of autophagy[455].

Chr14q24.3. The lead variant, rs59192843, is located within intron 6 of *BBOF1* (*basal body orientation factor 1*, also termed as *CCDC176*). There are no coding variants or eQTLs associated with the lead variant. Two variants in high LD (r^2<0.8), rs73301475 and rs17094157 scored highly in the integrated analysis of non-coding variants (Supplementary Table 8). These are located within an enhancer of *ENTPD5* (*ectonucleoside triphosphate diphosphohydrolase 5*) and the 3' UTR of *COQ6* (*coenzyme Q6, monooxygenase*), respectively. ENTPD5 hydrolyses UDP to UMP to promote protein N-glycosylation and folding. It has been shown that ENTPD5 was upregulated in cell lines and primary human tumour samples with active AKT, promoting cell growth and survival[456]. AKT activation also contributes to the elevation of aerobic glycolysis seen in tumour cells, known as the Warburg effect. Of note, ENTPD5 was also involved in stimulating glycolysis by providing substrates for cytidine monophosphate kinase-1 that converts UMP to UDP using a phosphate molecule generated during the ATP hydrolysis cycle[457]. COQ6 is an evolutionarily conserved monooxygenase, belonging to the ubiH/COQ6 family, which is required for the biosynthesis of coenzyme Q10 (or ubiquinone), an essential component of the mitochondrial electron transport chain and one of the most potent lipophilic antioxidants implicated in the protection of cell damage by reactive oxygen species. Gene-ablated mouse model showed abnormal embryo size and growth retardation (MGI: 5548683). Mutations in this gene are associated with autosomal recessive coenzyme Q10 deficiency-6, which manifests as nephrotic syndrome with sensorineural deafness[458].

Chr14q32.11. In this locus the variants are focused across *CALM1* (*calmodulin 1*). There is an eQTL co-localised with *CALM1* expression in testis. Two SNPs (rs12885713 and rs2300496) are within the *CALM1* promoter/enhancer region and predicted to have regulatory function. *CALM1* encodes a member of the EF-hand calcium-binding protein family, regulating a number of protein kinases and phosphatases, among which CP110, by interacting with CALM1 and centrin, regulates centrosome function and cytokinesis[459].

Chr14q32.33. The lead SNP, rs117536281, is located upstream of *CDCA4* (*cell division cycle associated 4*). *CDCA4* encodes a member of the E2F family of transcription factors, regulating spindle organization, cytokinesis and cell proliferation, which may be also involved in

differentiation of hematopoietic stem cells and progenitor cell lineage[460]. There are no coding variants or eQTL data for this locus.

Chr15q14. This locus consists of two associated SNPs, rs9972513 and rs12324579, which are located in an intergenic region upstream of both *c15orf53* and *RASGRP1* (*RAS guanyl releasing protein 1*). There are no coding variants or eQTL data for this locus. *C15orf53* is a protein coding gene with uncharacterised functions, with disputable evidence suggesting its implication with schizophrenia and bipolar disorder[461]. *RASGRP1* encodes a protein that functions as a calcium- and diacylglycerol (DAG)-regulated nucleotide exchange factor specifically activating Ras through the exchange of bound GDP for GTP. RASGRP1 contains a pair of calcium-binding EF hands and a DAG-binding domain[462]. The RASGRP1-mediated Ras activation regulates T cell proliferation, development and homeostasis[463].

Chr15q21.2. There are 17 SNPs clustered around the 5' end of *ATP8B4* (*ATPase phospholipid transporting 8B4 (putative)*). There are no coding variants or eQTL data for this locus. *ATP8B4* encodes a member of the cation transport ATPase (P-type) family and type IV subfamily, which consists of a P4-ATPase flippase complex that catalyses the hydrolysis of ATP coupled to phospholipid translocation across various membranes, playing a role in vesicle biosynthesis and lipid signalling transduction[464,465]. Deleterious rare variants within this gene have been associated with systemic sclerosis, for which the principal cause of death was pulmonary diseases, including interstitial lung disease and pulmonary arterial hypertension[466]. An intronic common variant at the distal promoter region of this gene has been reported to be associated with Alzheimer's Disease[467].

Chr15q21.3. This single variant, rs117610974 is located in an intergenic region, ~220kb downstream of the closest gene, *UNC13C* (*unc-13 homolog C*), which might be implicated with vesicle formation during exocytosis, with potential capabilities of diacylglycerol and calcium binding[468]. However, there is no evidence to suggest what role this lead variant may have.

Chr15q22.31. The lead variant, rs55710439, is located within intron 6 of *ANKDD1A* (*ankyrin repeat and death domain containing 1A*). There is an eQTL for this gene co-localised in one tissue. Little is known about the ANKDD1A protein, except that it contains an ankyrin repeat domain and a death domain, both of which function in the protein-protein interaction. A closely-related SNP (in LD, r^2<0.8), rs57438358, predicted to have potential functional effects, is located within the 3'UTR of *SPG21* (*SPG21, maspardin*), a gene which is mutated in mast syndrome.

Chr16p13.3. This is a single variant, rs11640926, located within intron 5 on *CACNA1H*. There is no supporting evidence to suggest the effects of this variant. *CACNA1H* encodes a protein component of the voltage-dependent calcium channel complex, a T-type calcium channel that belongs to the "low-voltage activated" group, which plays an essential role in both central neurons and cardiac nodal cells and supports calcium signalling in secretory cells and vascular smooth muscle[469,470]. It is associated with a form of familial hyperaldosteronism, clinically characterised by hypertension, elevated aldosterone levels and abnormal adrenal steroid production[471]; and another genetic rare disease, the Childhood Absence Epilepsy 6[472].

Chr16q22.1. The most significantly associated variants in this region are located within and around *TERF2* (*telomeric repeat binding factor 2*), a component of the shelterin complex. TERF2 protein directly and specifically binds to the telomeric double-stranded repeats, and by interacting with other telomeric factors forming a T-loop configuration that protects chromosome ends from disruptive end-to-end joining and ligation to exogenous DNA. Mutant forms of this gene induced DNA fusion, such as formation of anaphase bridges and lagging or ring-like chromosomes[473,474].
There is evidence that the variants affect expression of several genes in this region, with the strongest evidence for *TERF2* (S-PrediXcan and COLOC, Supplementary Table 7). Longer LTL is associated with reduced expression of *TERF2*, consistent with TERF2 being a negative regulator of telomere length[475]. One variant predicted to have a functional effect, rs9939705, is located within an enhancer region upstream of *TERF2*. There is also evidence to suggest that expression of two other genes (*COG8*, and *PDF*) are also affected by the associated variants.

Chr16q23.1. Variants at this locus show co-localisation with eQTLs for *RFWD3* (*ring finger and WD repeat domain 3*) in multiple tissues. RFWD3 is a ubiquitin ligase that interacts with and ubiquitinates replication protein A (RPA), which has been shown to be essential for DNA replication and repair. Upon replication stress, RPA was recruited to stalled replication folks and ubiquitinated by the RFWD3, an essential process for recovery and homologous recombination-mediated DNA repair[476]. RFWD3 also ubiquitinates and stabilises p53/TP53 in response to DNA damage, thereby regulating the cell cycle checkpoint[477]. This gene was also clinically attributable to the Fanconi anaemia, an autosomal recessive inheritance disease manifested with chromosomal instability, bone marrow failure, dermal pigmentary changes and predisposition to malignancies (OMIM #614151).

Chr16q23.3. The association signal at this locus is across *MPHOSPH6* (*M-phase phosphoprotein 6*). There is strong eQTL evidence (S-PrediXcan and COLOC) in multiple tissues to support the associated variants influencing MPHOSPH6 expression. MPHOSPH6 is a component of the RNA exosome, a protein complex required for the degradation of RNA molecules and is required for the 3' processing of the 5.8S rRNA[272]. There is also evidence that MPHOSPH6 interacts with PARN (poly(A)-specific ribonuclease)[478], an important regulator of mRNA catabolism which is also required for the formation of mature *TERC* RNA[269].

Chr17q25.3. The lead variant (rs144204502) is situated within the 5' UTR of *TK* (*thymidine kinase 1*), with evidence of regulatory functions (Supplementary Table 8). There are co-localised eQTLs for *TK1* in three tissues. *TK1* encodes a cytosolic enzyme that catalyses the conversion of thymidine to dTMP, which is the first step of the salvage pathway of dTTP biosynthesis, essential for DNA replication. There are two forms of the TK enzyme, besides the TK1, TK2 catalyses the same reaction but in the mitochondria. The activity of TK1 is delicately regulated by a configurational transition, changing from dimer to tetramer upon increases in ATP and enzyme concentrations, with a consequently accompanied upregulation of catalytic efficiency[274]. This regulatory fine-tuning of TK1 activity ensured a balanced pool of nucleic acid precursors. High TK1 expression was detected in numerous types of cancers, including gastrointestinal adenocarcinomas and oesophageal and uterine squamous cell carcinomas[479].

Chr18p11.32. All variants within the locus are located within the *TYMS* (*thymidylate synthetase*) gene, either within the intronic or the 3'UTR regions. There is an eQTL for *TYMS* co-localised in one tissue. TYMS is involved in the *de novo* biosynthesis of dTMP, catalysing the methylation of dUMP to dTMP using a serine-derived one-carbon donor, the 5,10-methyleneTHF[278]. *TYMS* has been targeted for cancer chemotherapeutics, as high expression of which has been detected in various types of cancers, including gastrointestinal adenocarcinomas and squamous cell uterine carcinomas[479].

Chr19p13.3. The lead variant is located within intron 5 of *NMRK2* (*Nicotinamide Riboside kinase 2*), with 6 SNPs in high LD (r^2<0.8) located around this gene. NMRK2 enzyme catalyses the phosphorylation of nicotinamide riboside (NR) and nicotinic acid riboside (NaR) to form nicotinamide mononucleotide (NMN) and nicotinic acid mononucleotide (NaMN), the vitamin precursors of NAD^+, which is required for the function of Sirtuins, a key player in lifespan extension and energy metabolism[480]. It has been demonstrated that increased NAD^+ biosynthesis elevated the Sirtuin 2 function, which improved the subtelomeric gene silencing effects and elongated replicative lifespan in eukaryotic cell models[480]. One further variant in high LD, located upstream of *DAPK3* (*death associated protein kinase 3*), is a regulator of apoptosis. There is no functional data supporting any gene candidates at this locus.

Chr19p12. The lead variant is intergenic, located between *ZNF257* and *ZNF208*, with closer proximity to the former. There is eQTL evidence for both *ZNF257* and *ZNF265*, yet stronger for the *ZNF257* (Supplementary Table 7). *ZNF257* encodes a member of a zinc finger protein family, the Krüppel-like zinc finger subfamily, signified by a consensus sequence of TGEKPYX (X denotes any amino acids) between concatenated zinc finger motifs[481]. The proteins have the KRAB domain at their amino terminus, which determines the specificity of binding to DNA and other transcriptional co-regulators.

Chr19q13.2. The single associated variant, rs11665818, is located within an intergenic region, downstream of *INFL2* (*interferon lambda 2*, also termed *IL28a*) and within a cytokine gene cluster that consists of three closely related *INFL* genes. *INFL2* encodes a protein with antiviral activities, predominantly in the epithelial tissues[482]. There is no supporting functional evidence at this locus.

Chr20p12.3a. The lead and one variant in high LD (r^2<0.8) are located upstream of *PROKR2* (*Prokineticin receptor 2*), a G protein-coupled receptor for the prokineticin 2, which is a secreted protein expressed in gut and brain, and has been shown to oscillate on a circadian basis[483]. Homozygous gene-knockout mice showed impaired circadian behaviour and thermoregulation (MGI:2181363). Mutations in this gene led to gonadotropin-releasing hormone deficiency and hypogonadism[484]. There are no coding variants or eQTLs associated with this locus.

Chr20p12.3b All variants of this locus are located within an intergenic region, with the closest gene being LINC01706 (long intergenic non-coding RNA 1706), an uncharacterised non-coding transcript.

Chr20q11.23. The association signal spans two genes *MROH8* (*maestro heat like repeat family member 8*) and *RBL1* (*RB transcriptional corepressor like 1*). There is eQTL evidence to support

changes in both *RBL1* and *SAMHD1* (*SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1*) expression. RBL1 functions as a transcriptional repressor for E2F binding sites-containing genes[485], which shares similarity in amino acid sequence and biochemical features to the *retinoblastoma 1* (*RB1*) gene product that functions as a tumour suppressor implicated in cell cycle regulation. *SAMHD1* encodes a dNTP triphosphohydrolase (dNTPase) that converts deoxynucleoside triphosphates (dNTPs) to deoxynucleosides. The gene expression was regulated during cell cycle to maintain a homeostatic pool of dNTP, required for DNA replication [250]. Studies have suggested an antiretroviral role of SAMHD1 in dendritic and myeloid cells by depleting the intracellular pool of dNTPs[486,487].

Chr20q13.33. There are four independent signals within this locus, which harbours several genes, including the DNA helicase *RTEL1* (*regulator of telomere elongation helicase 1*). There are non-synonymous coding variants in *RTEL1* and *ZBTB46* (*zinc finger and BTB domain containing 46*) although neither are predicted to be deleterious. There are eQTLs for *RTEL1*, *STMN3*(*stathmin 3*) and *TNFRSF6B* (*TNF receptor superfamily member 6b*, also termed *decoy receptor 3*). *RTEL1* encodes an ATP-dependent DNA helicase that functions in the regulation of telomeres, DNA repair and genomic integrity. RTEL1 facilitates access of telomerase to the 3' ends of telomeres by transiently dismantling the T-loop configuration, a lariat-like structure that protects telomeres from degradation and deleterious DNA damage response[488]. Mutations of this gene led to Hoyeraal Hreidarsson syndrome, a clinically severe form of dyskeratosis congenita, of which half of the inherited families carry germline mutations of telomere-related genes[489]. Loss-of-function missense variants of this gene was found to be associated with idiopathic pulmonary fibrosis and shortened telomere lengths[490]. *STMN3* gene encodes a member of the stathmin protein family, which shows microtubule-destabilizing activity and is known to be involved in the development of central nervous system and glioma pathology[491]. *TNFRSF6B* is a regulator of apoptosis and has been linked to angiogenesis[492–494]. *ZBTB46* gene encodes a member of a large BTB zinc-finger protein family, characterised by a DNA binding motif that consists of a tandem array of C2H2 krüppel-like zinc fingers at the carboxyl terminus, with each finger containing a consensus sequence of ~30 amino acids and an embedded zinc ion[495]. In contrast, the BTB domains at the amino termini are more divergent across the family, mainly contributing to the hetero- or homo-dimerization. The BTB domain determines DNA binding specificity and recruitment of co-regulators to form higher chromosomal structures[495]. ZBTB46 has been shown to function as a transcriptional repressor involved in prostate cancer malignancy and cell cycle regulation[496]. Recently, studies have identified another member of the BTB zinc-finger protein family, ZBTB48, also termed as the telomeric zinc finger–associated protein, to be specifically associated with telomeres via the zinc finger domain. Further investigation demonstrated that it was preferentially bound to longer telomeres where protein components of the shelterin complex are rather sparse[497]. Experimental studies suggested that the ZBTB48 might compete with the TERF2 for binding to the telomeric DNA repeats, thereby setting an upper limit of the telomere length, which can further influence lifespan and cancer susceptibility[497,498]. Because the zinc finger domain is conserved among all members of the family, we speculated that the ZBTB46 was also capable of binding to the telomeric DNA, regulating telomere homeostasis via similar mechanisms. However, further experiments are required to validate this hypothesis.

Chr21q22.3. The lead variant is a loss-of-stop mutation in *KRTAP10-4* (*keratin associated protein 10-4*), which was located within a cluster of related genes, encoding proteins that form disulfide bonds between cysteine residues in hair keratins. A genome-wide siRNA-based screen implicated this gene with the homologous recombination DNA double-strand break repair[499]. Although transcripts lacking stop codons would be targeted for degradation, there is no eQTL evidence to suggest loss of expression with this allele, possibly due to poor detection of this transcript in GTex (Median transcripts per million=0). There is one variant in high LD, located within intron 2 of *TSPEAR* (*thrombospondin type laminin G domain and EAR repeats*), a regulator of the NOTCH signalling.

Chr22q13.31. This is a single variant located within intron 1 of *KIA1644* (Also termed *SHISAL1*). There is no supporting functional data for gene prioritisation at this locus.

Systematic literature review on longitudinal changes of TL
Searching strategies applied:

	Query	Number of items
Telomeres #1	telomere[Mesh] OR telomeres[ti] OR telomere[ti]	15,044
Changing rates #2	rate[tiab] OR rates[tiab] OR shortening[tiab] OR abrasion[tiab] OR attrition[tiab] OR erosion[tiab] OR extension[tiab] OR acceleration[tiab] OR accelerating[tiab] OR lengthening[tiab] OR elongation[tiab] NOT "telomere elongation helicase1"[tiab] NOT "alternative lengthening of telomeres"[tiab]	2,619,963
Cohort studies #3	"cohort studies"[Mesh] OR cohort[tiab] OR cohorts[tiab] OR "longitudinal studies"[Mesh] OR longitudinal[tiab] OR "long term"[tiab] OR "short term"[tiab] OR prospective[tiab] OR retrospective[tiab]	2,810,034
Genetics #4	genetics[Mesh] OR genetic[tiab] OR gene[tiab] OR genes[tiab] OR genome[tiab] OR genomes[tiab] OR	2,438,596
Combined strategy	#1 AND #2 AND (#3 OR #4)	2,043

Summary of the study results:
Abbreviations: SCDS: Seychelles Child Development Study, HBCS: Helsinki Birth Cohort Study, HSS: Heart and Soul Study, BEIP: Bucharest Early Intervention Project, SATSA: Swedish Adoption/Twin Study of Aging, GEMINAKAR: Genes, Familiar and Common Environment for the Development of Insulin Resistance, Abdominal Adiposity, and Cardiovascular Risk Factors, CBMC: Cord blood mononuclear cell, PBMC: peripheral blood mononuclear cell, NESDA: Netherlands Study of Depression and Anxiety, LBC: Lothian Birth Cohort, DMHDS: Dunedin Multidisciplinary Health and Development Study, ESTHER: Epidemiological investigations on chances of preventing, recognizing early and optimally treating chronic diseases in an elderly population, PREDIMED-NAVARRA: PREvención con DIeta MEDiterránea-NAVARRA, PREVEND: Prevention of REnal and Vascular ENd stage Disease, CHS: Cardiovascular Health Study, JLRC: Jerusalem Lipid Research Clinic, MRC-NSHD: MRC-National Survey of Health and Development, HAS: Hertfordshire Ageing Study, CCHS: Copenhagen City Heart Study, HBLS: Harvard Boilermakers Longitudinal Study, ERA: Evolution de la Rigidite Arterielle, BHS: Bogalusa Heart Study, CaPS: Caerphilly Prospective Study, LSADT: Longitudinal Study of Aging Danish Twins, MHAS: MacArthur Health Aging Study, NSHDS: North Sweden Health and Disease Study.

Index	Publication	Study cohort	Time Interval	Study Participants
			(at maximum)	*(baseline age in years)*
1	Yeates AJ. J Nutr. 2017	SCDS	5 years	newborn babies (0)
2	Dowd JB. J. Infect. Dis. 2017	Whitehall II	3 years	healthy individuals (53-76)
3	Ventura Ferreira MS. Ann Hematol. 2017		1 year	patients
4	Vasu V. Plos One. 2017		5 years	preterm infants (0)
5	Toupance S. Hypertension. 2017		10 years	French (31-76)
6	Steptoe A. J Clin Endocrinol Metab. 2017		3 years	healthy individuals (53-76)
7	Eriksson JG. Am J Clin Nutr. 2017	HBCS	10 years	healthy individuals (71)
8	de Melo AS. Gene. 2017		5 years	healthy women (30)
9	Goglin SE. PLoS One. 2016	HSS	5 years	patients with stable CAD
10	Humphreys KL. Psychiatry Res. 2016	BEIP	6 years	children (8)
11	See VH. Prostaglandins Leukot Essent Fatty Acids. 2016		12 years	offspring (0)
12	Ping F. J Diabetes Investig. 2017		6 years	T2D patients
13	Berglund K. Aging (Albany NY). 2016	SATSA	20 years	twins (69)
14	Townsley DM. N Engl J Med. 2016		2 years	patients with telomere diseases
15	Verhulst S. Diabetologia. 2016	GEMINAKAR	12 years	twins (37)
16	Lin J. J Immunol Res. 2016		18 months	healthy premenopausal women (41.9)
17	Wojcicki JM. Mol Genet Genomics. 2016		1 year	mother-child pairs
18	Kato S. Blood Purif. 2016		1 year	Japanese incident dialysis patients
19	Verhoeven JE. Am J Psychiatry. 2016	NESDA	6 years	patients with depressive/anxiety disorders
20	Révész D. Psychoneuroendocrinology. 2016	NESDA	6 years	patients with depressive/anxiety disorders and healthy controls (18-65)
21	Harris SE. Mech Ageing Dev. 2016	LBC1936/1921	6/13 years	healthy individuals (70/79)
22	Thomson WM. J Clin Periodontol. 2016	DMHDS	12 years	healthy individuals (26)
23	Müezzinler A. Exp Gerontol. 2016	ESTHER	8 years	healthy individuals (50-75)
24	Jenkins EC. Am J Med Genet B Neuropsychiatr Genet. 2016		3 years	DS patients with syndromes of cognitive impairment
25	Dalgård C. Int J Epidemiol. 2015	GEMINAKAR	12 years	twins (18-59)
26	García-Calzón S. Am J Clin Nutr. 2015	PREDIMED-NAVARRA	5 years	individuals at high cardiovascular disease risk (67)
27	Guzzardi MA. Ann Med. 2015	HBCS	10 years	healthy individuals (71)
28	Hou L. EBioMedicine. 2015	NAS	3 years	prevalent and incident cancer cases and others
29	Müezzinler A. Exp Gerontol. 2015	ESTHER	8 years	healthy individuals (50-75)
30	van Ockenburg SL. Psychol Med. 2015	PREVEND	6 years	healthy individuals (53)
31	Révész D. J Clin Endocrinol Metab. 2015	NESDA	6 years	patients with depressive/anxiety disorders and healthy controls (18-65)
32	Soares-Miranda L. Med Sci Sports Exerc. 2015	CHS	5 years	healthy individuals (73)
33	Cohen-Manheim I. Eur J Epidemiol. 2016	JRC	13 years	healthy individuals (28-32)

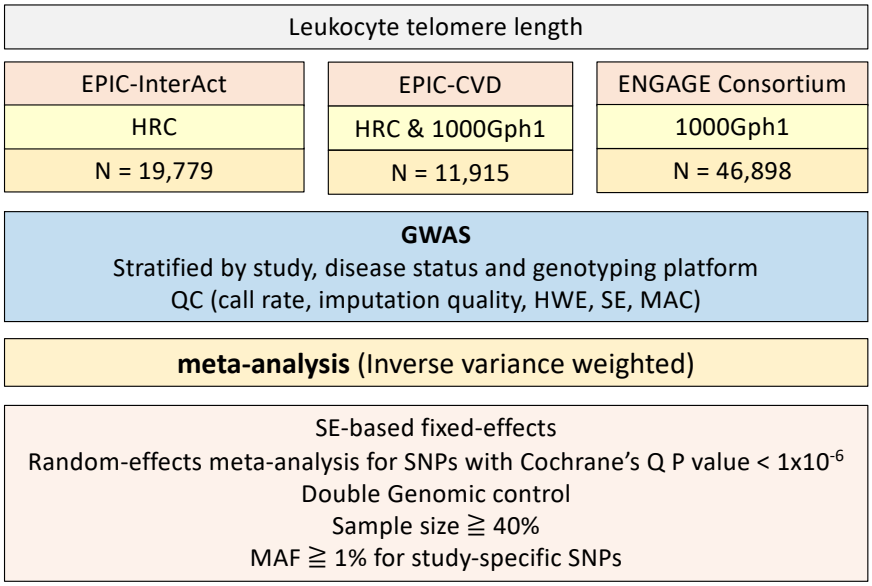
34	Ashbridge B. Biol Blood Marrow Transplant. 2015		1 year	patients with high-risk hematologic malignancies
35	Hjelmborg JB. J Med Genet. 2015	GEMINAKAR	12 years	twins (19-64)
36	García-Calzón S. Circ Cardiovasc Genet. 2015	PREDIMED-NAVARRA	5 years	individuals at high cardiovascular disease risk (55-80)
37	Puterman E. Mol Psychiatry. 2015		1 year	postmenopausal, non-smoking, disease-free women (50-65)
38	Masi S. Eur Heart J. 2014	MRC-NSHD	10 years	study participants (53)
39	Baylis D. Calcif Tissue Int. 2014	Normative Aging	10 years	study participants (67)
40	Tamayo M. Mutat Res. 2014		3 years	patients with ankylosing spondylitis/psoriatic arthritis
41	Rewak M. Biol Psychol. 2014	EdHealth	41 years	study participants (0)
42	Duggan C. J Natl Cancer Inst. 2014		30 months	patients with breast cancers
43	Weischer M. PLoS Genet. 2014	CCHS	10 years	study participants (20-100)
44	Wong JY. Genet Epidemiol. 2014	HBLS	29 months	boilermakers (43)
45	Bendix L. J Gerontol A Biol Sci Med Sci. 2014	Danish MONICA	10 years	study participants (30-70)
46	Huzen J. J Intern Med. 2014	PREVEND	6.6 years	study participants (39-60)
47	Verhulst S. Eur J Epidemiol. 2013	JLRC/BHS/ERA	13/12/9 years	study participants (30/31/58)
48	Gardner MP. PLoS One. 2013	CaPS/HAS/LBC1921/MRC-NHSD	8/9/7/9 years	study participants (65/67/79/53)
49	van Ockenburg SL. Psychol Med. 2014	PREVEND	6 years	study participants (53)
50	Benetos A. Ageing Cell 2013	JLRC/BHS/ERA/LSADT	13/12.4/9.5/10.8 years	study participants (30/31/58/75)
51	Steenstrup T. Eur J. Epidemiol. 2013	LSADT	10 years	study participants (73-81)
52	Bansal N. Am J Nephrol. 2012	HSS	5 years	patients with stable CAD
53	Biegler KA. Cancer Prev Res (Phila). 2012		4 months	patients with cervical cancers
54	Lobetti-Bodoni C. Mech Ageing Dev. 2012		22 months	patients with chronic myeloid leukemia (23-88)
55	Kark JD. Am J Clin Nutr. 2012	JLRC	13 years	study participants (30)
56	Selleri S. J Allergy Clin Immunol. 2011		9 years	patients with immune deficiency (1-5)
57	McCracken J. Environ Health Perspect. 2010	NAS	7 years	never-smoking men (56-94)
58	Shlush LI. Mech Ageing Dev. 2011		1 year	CCORDA diver group (19)
59	Chen W. J Gerontol A Biol Sci Med Sci. 2011	BHS	12.4 years	study participants (31)
60	Farzaneh-Far R. JAMA. 2010	HSS	5 years	patients with stable CAD
61	Farzaneh-Far R. PLoS One. 2010	HSS	5 years	patients with stable CAD
62	Aviv A. Am J Epidemiol. 2009	BHS	12.4 years	study participants (white 31.4/black 37.4)
63	Epel ES. Aging (Albany NY). 2008	MHAS	2.5 years	Caucasian participants (70-79)
64	Nordfjäll K. PLoS Genet. 2009	NSHDS	10 years	participants from a multigenerational cohort
65	Ehrlenbach S. Int J. Epidemiol. 2009	Bruneck	10 years	study participants (60)

Index	Sample Size	Measurement	Tissue or Cell Line	Risk Factors or Consequences Tested
1	229	qPCR	leucocytes	PUFA status and methylmercury levels
2	400	qPCR	leucocytes	human herpesviruses
3	49	qPCR	leucocytes	acute myeloid leukemia
4	5	qPCR	leucocytes	gestational age
5	154	TRF-Southern blots	leucocytes	carotid atherosclerotic plaques
6	411	qPCR	leucocytes	cortisol responses
7	812	qPCR	leucocytes	serum phenylalanine concentration
8	165	qPCR	leucocytes	birth weight
9	608	qPCR	leucocytes	mortality
10	79	MMP-qPCR	saliva	institutional care
11	98	qPCR	CBMCs and PBMCs	LCPUFA supplementation during pregnancy
12	70	qPCR	leucocytes	NAFLD
13	636	qPCR	leucocytes	age
14	27	qPCR	leucocytes	Danazol Treatment
15	338	TRF-Southern blots	leucocytes	insulin resistance
16	39	qPCR	PBMCs and lymphocytes	circulating immune cell types
17	70	qPCR	leucocytes	baseline telomeres and cross-generational change
18	59	qPCR	leucocytes	baseline telomeres and leukocyte counts
19	2,292	qPCR	leucocytes	depressive and anxiety disorders
20	2,981	qPCR	leucocytes	baseline telomeres, lifestyle factors and diseases
21	1091/550	qPCR	leucocytes	cognitive and physical abilities
22	661	qPCR	leucocytes	periodontitis
23	961	qPCR/TRF (n=20)	leucocytes	BMI-related variables
24	5	FITC-labeled PNA probes	T-lymphocytes	clinical progression of AD for patients with DS
25	734	TRF-Southern blots	leucocytes	menopausal status and age
26	520	qPCR	leucocytes	diet-associated inflammation
27	1,082	qPCR	leucocytes	cardiometabolic risk factors
28	792	qPCR	leucocytes	cancer development
29	961	qPCR/TRF (n=20)	leucocytes	smoking
30	1,094	MMP-qPCR	leucocytes	psychosocial stress
31	1,808	qPCR	leucocytes	cardiometabolic risk factors
32	582	TRF-Southern blots	leucocytes	physical activity and fitness
33	497	TRF-Southern blots	leucocytes	cognitive function

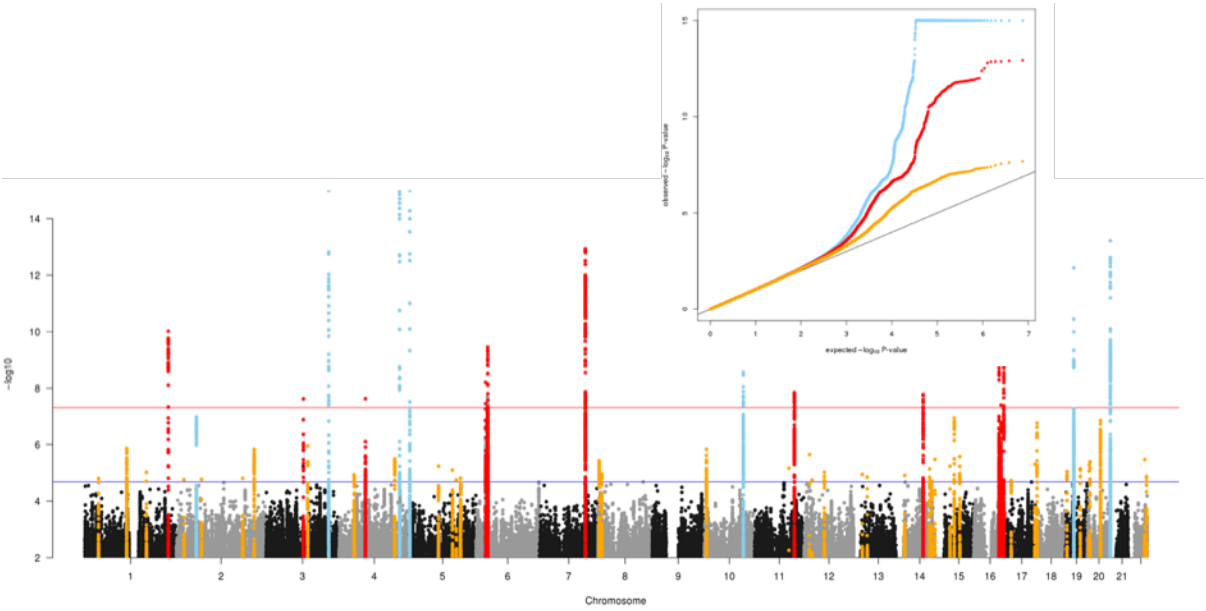
34	13	TRF-Southern blots	CBMCs and PBMCs	cord blood transplantation
35	652	TRF-Southern blots	leucocytes	heritability
36	521	qPCR	leucocytes	PPARγ2 polymorphysm and diet intervention
37	239	qPCR	leucocytes	life stressors and health behaviours
38	1,033	qPCR	leucocytes	cardiovascular risk factors
39	253	qPCR	leucocytes	low-grade systemic inflammation and grip strength
40	44/42	qPCR	leucocytes	spondyloarthritis
41	143	qPCR	leucocytes	race
42	478	qPCR	leucocytes	all-cause or breast cancer-specific mortality
43	4,576	qPCR	leucocytes	lifestyle factors morbidity and mortality
44	87	qPCR	leucocytes	LINE-1 and Alu methylation
45	1,356	qPCR	leucocytes	lifestyle factors, age and mortality
46	8,074	qPCR	leucocytes	metabolic traits and smoking
47	620/271/185	TRF-Southern blots	leucocytes	RTM effect and baseline telomere
48	966/656/493/2558	qPCR	leucocytes	physical performance
49	3,432	MMP-qPCR	leucocytes	neuroticism
50	620/271/185/80	TRF-Southern blots	leucocytes	telomere length at censor
51	80	TRF-Southern blots	leucocytes	age
52	608	qPCR	leucocytes	kidney function
53	22	flow-FISH	PMBCs	chronic stress response
54	59	TRF-Southern blots	granulocytes, and PMBCs	hematopoiesis upon treatment of chronic myeloid leukemia
55	609	TRF-Southern blots	leucocytes	energy intake and macronutrients
56	12	qPCR	BM/PB compartments and T cells	hematopoietic stem cell gene therapy
57	165	qPCR	leucocytes	ambient air pollution
58	14	flow-FISH/TRF	granulocytes and lymphocytes	hyperbaric oxidative stress
59	271	TRF-Southern blots	leucocytes	age
60	608	qPCR	leucocytes	blood levels of marine omega-3 fatty acids
61	608	qPCR	leucocytes	baseline telomeres and cardiometabolic risk factors
62	635	TRF-Southern blots	leucocytes	baseline telomeres and lifestyle factors
63	236	qPCR	leucocytes	mortality
64	959	qPCR	leucocytes	baseline telomeres and tumor development
65	510	qPCR/TRF (n=56)	leucocytes	basline telomeres, lifestyle factors and mortality

Supplementary Figures

Supplementary Figure 1. Study design. Schematic graph to illustrate study design of the LTL GWAS meta-analysis. GWAS was conducted in each individual study cohort, stratified by genotyping platform and disease status. SNP genotyping, GWAS and meta-analyses as well as the corresponding QC procedures were described in detail in section 2.2.3 and 2.2.4.

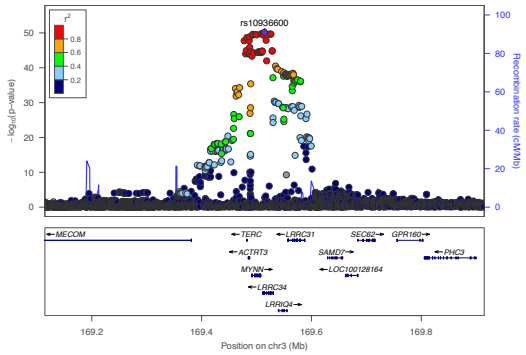


Supplementary Figure 2. Manhattan Plot. Manhattan plot with quantile-quantile plot inlay. Known loci were labelled in blue, novel loci associated with LTL at genome-wide significance (*p*-value<$5x10^{-8}$, red line) in red, and at FDR threshold of 5% (blue line) in orange.

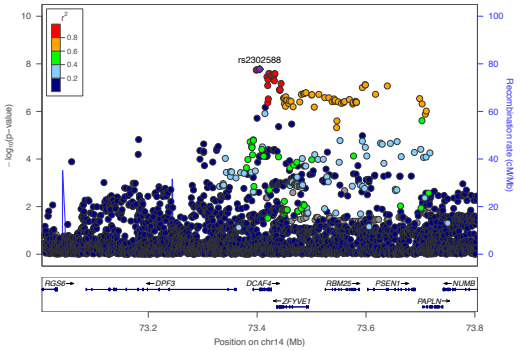


Supplementary Figure 3: Regional plots of genome-wide significant loci (regions around conditionally independent lead variants). Regional plots of genome-wide significant loci (400kb windows encompassing conditionally independent variants, except the *TERT* locus which is illustrated as a 200kb window).

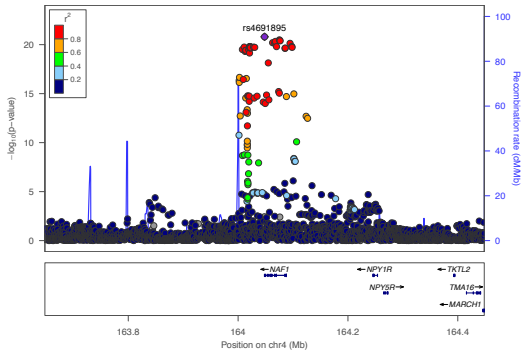
TERC chr3:169314585-169714585



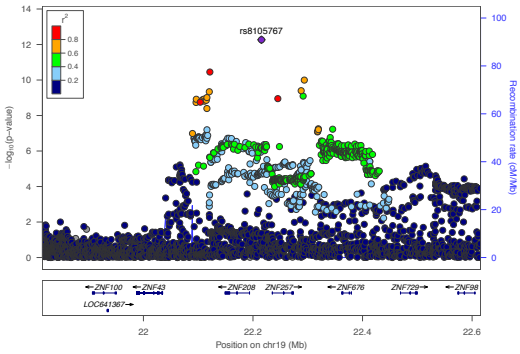
DCAF4 chr14:73204752-73604752



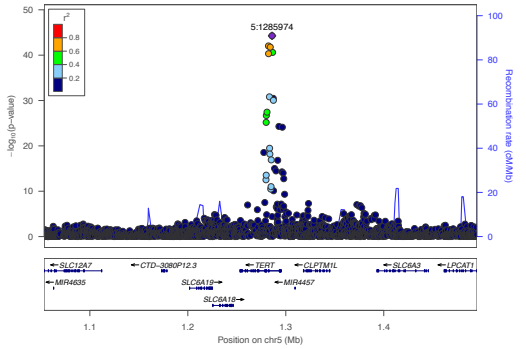
NAF1 chr4:163848199-164248199



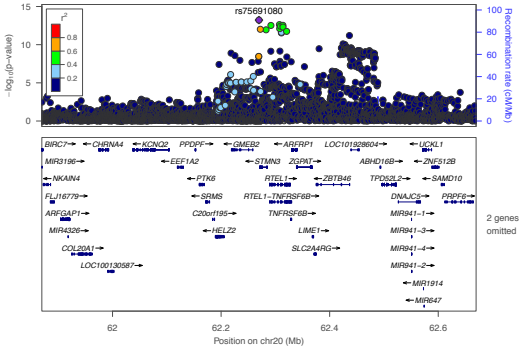
ZNF208 chr19:22015441-22415441



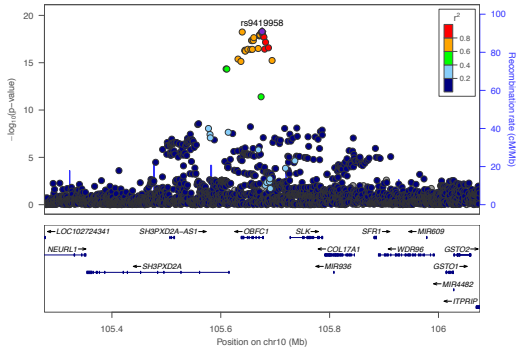
TERT chr5:1185974-1385974



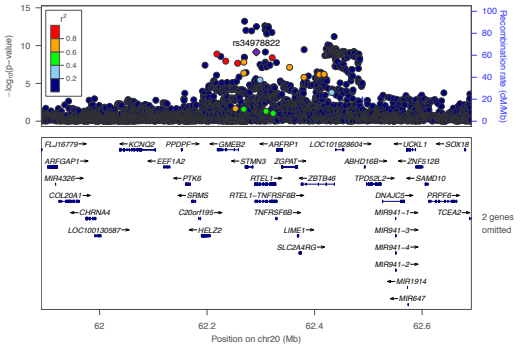
RTEL1/STMN3 chr20:62069750-62469750



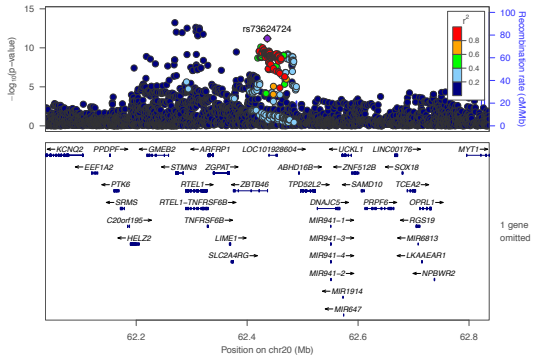
OBFC1 chr10:105475946-105875946



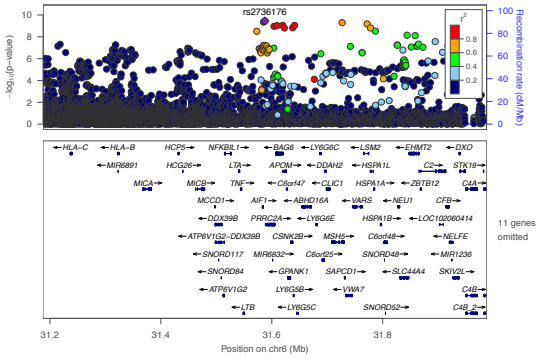
RTEL1 chr20:62091599-62491599



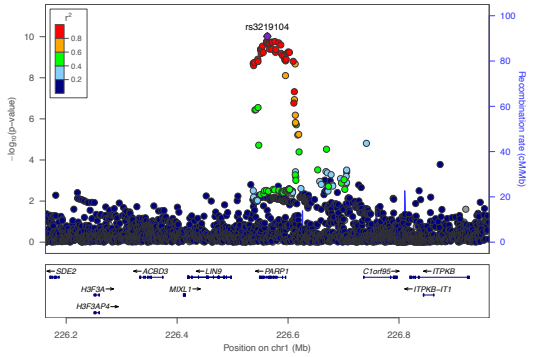
RTEL1/ZBTB46 chr20:62236398-62636398



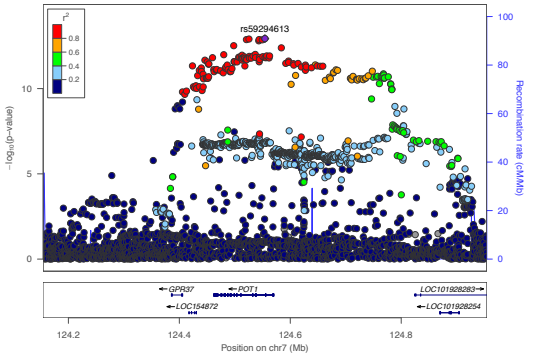
PRRC2A chr6:31387561-31787561



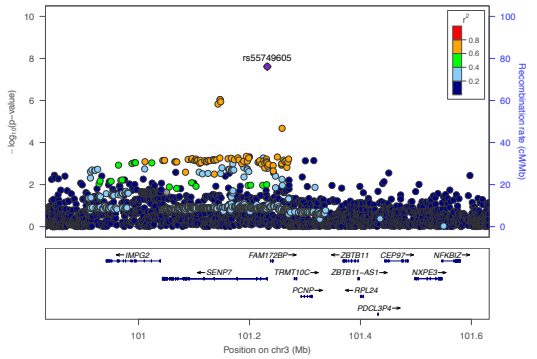
PARP1 chr1:226362621-226762621



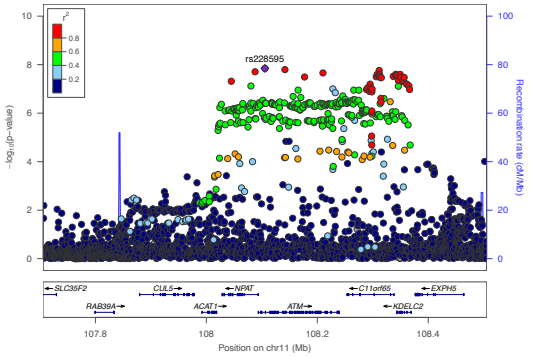
POT1 chr7:124354267-124754267



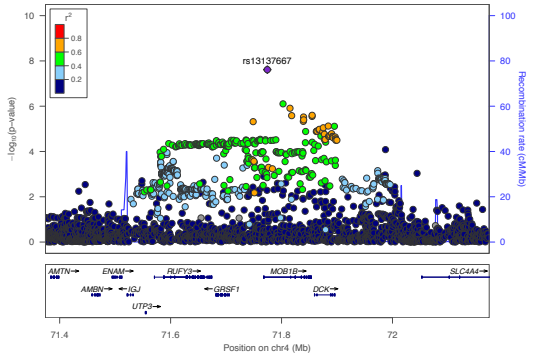
SENP7 chr3:101032093-101432093



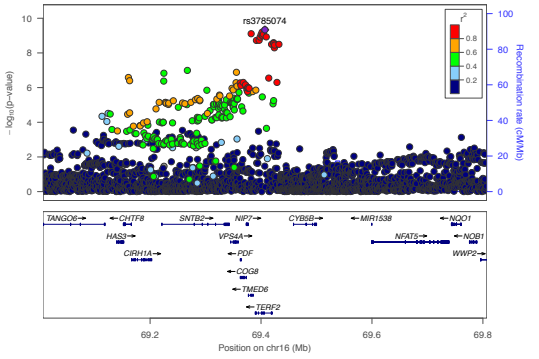
ATM chr11:107905593-108305593



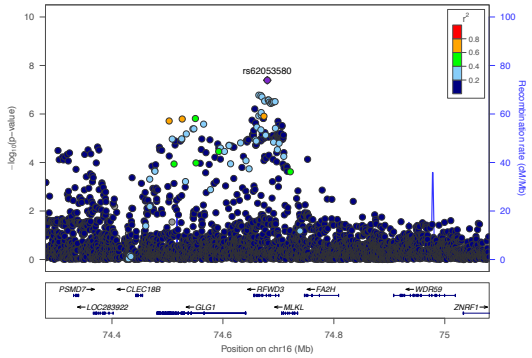
MOB1B chr4:71574347-71974347



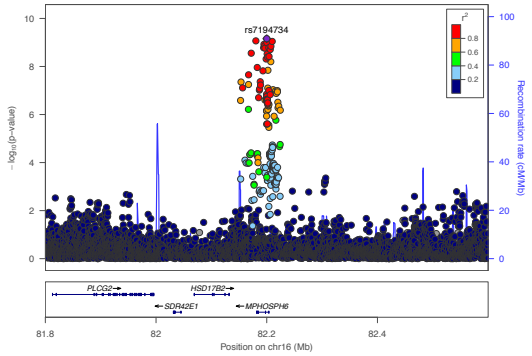
TERF2 chr16:69206986-69606986



RFWD3 chr16:74480074-74880074

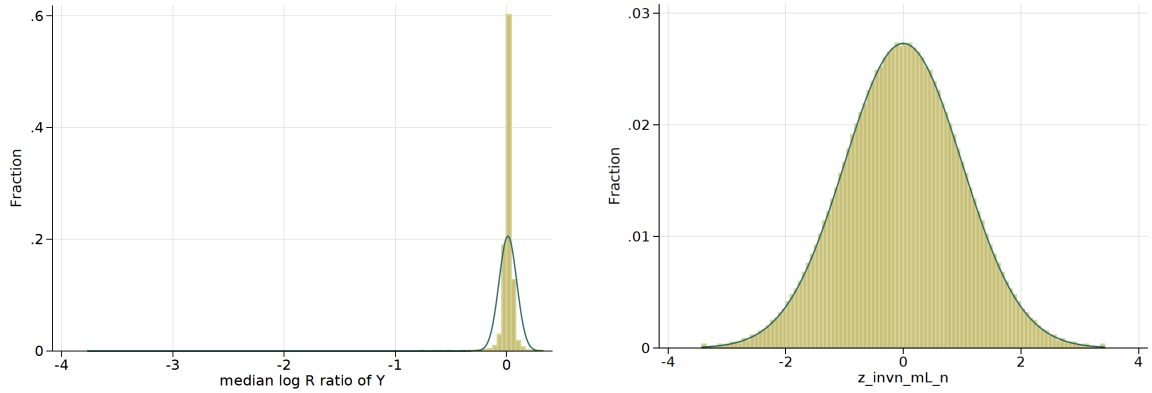


MPHOSPH6 chr16:81999980-82399980

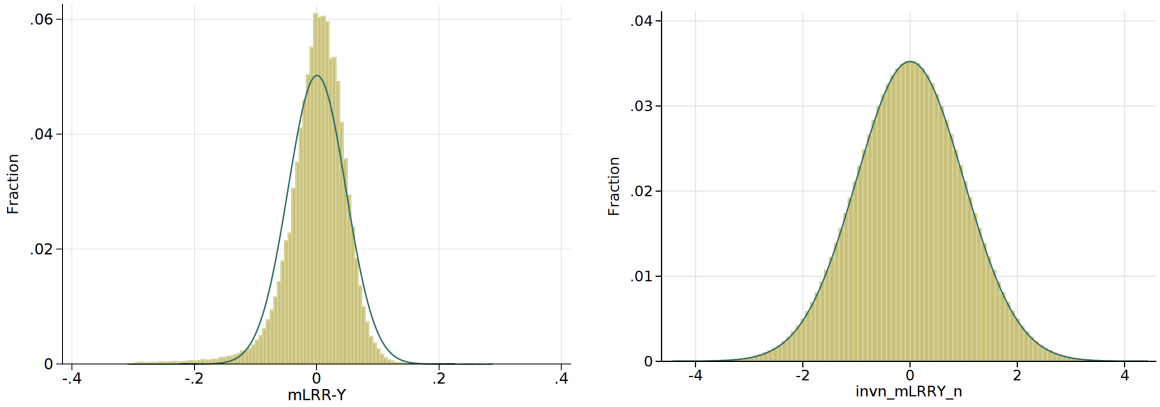


Supplementary Figure 4: Distributions of mLRRY values in **A.** EPIC-InterAct and **B**. UK biobank, before (left) and after (right) data transformation. Z_invn_mL_n means the standardised values of mLRRY after a series of data transformation (winsorisation at 5SD, followed by inverse normal transformation and z-standardisation).

A. EPIC-InterAct

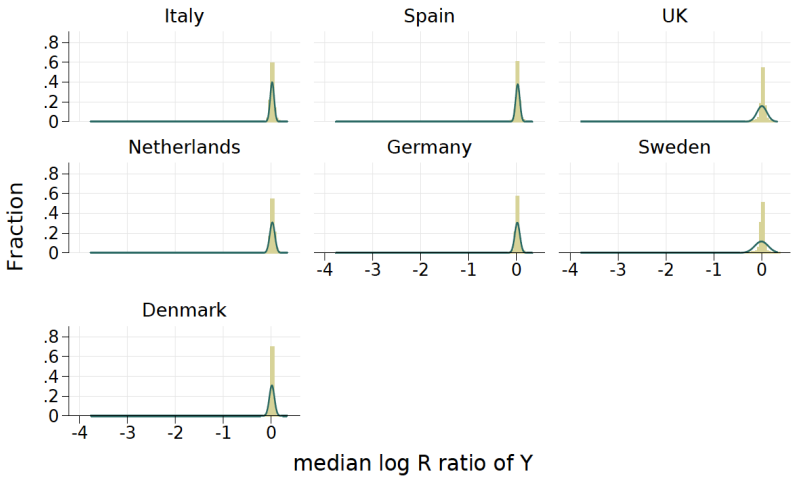


B. UK biobank

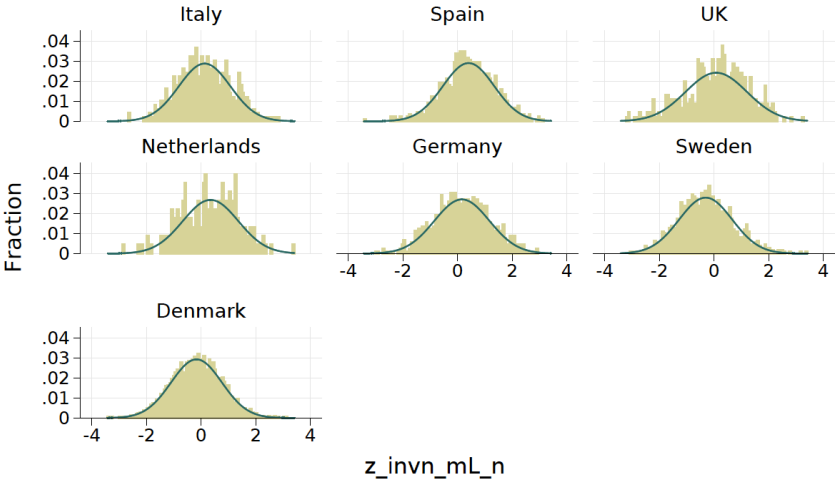


Supplementary Figure 5: Distribution of mLRRY values in each EPIC-InterAct participating country separately, before (upper) and after (bottom) data transformation.

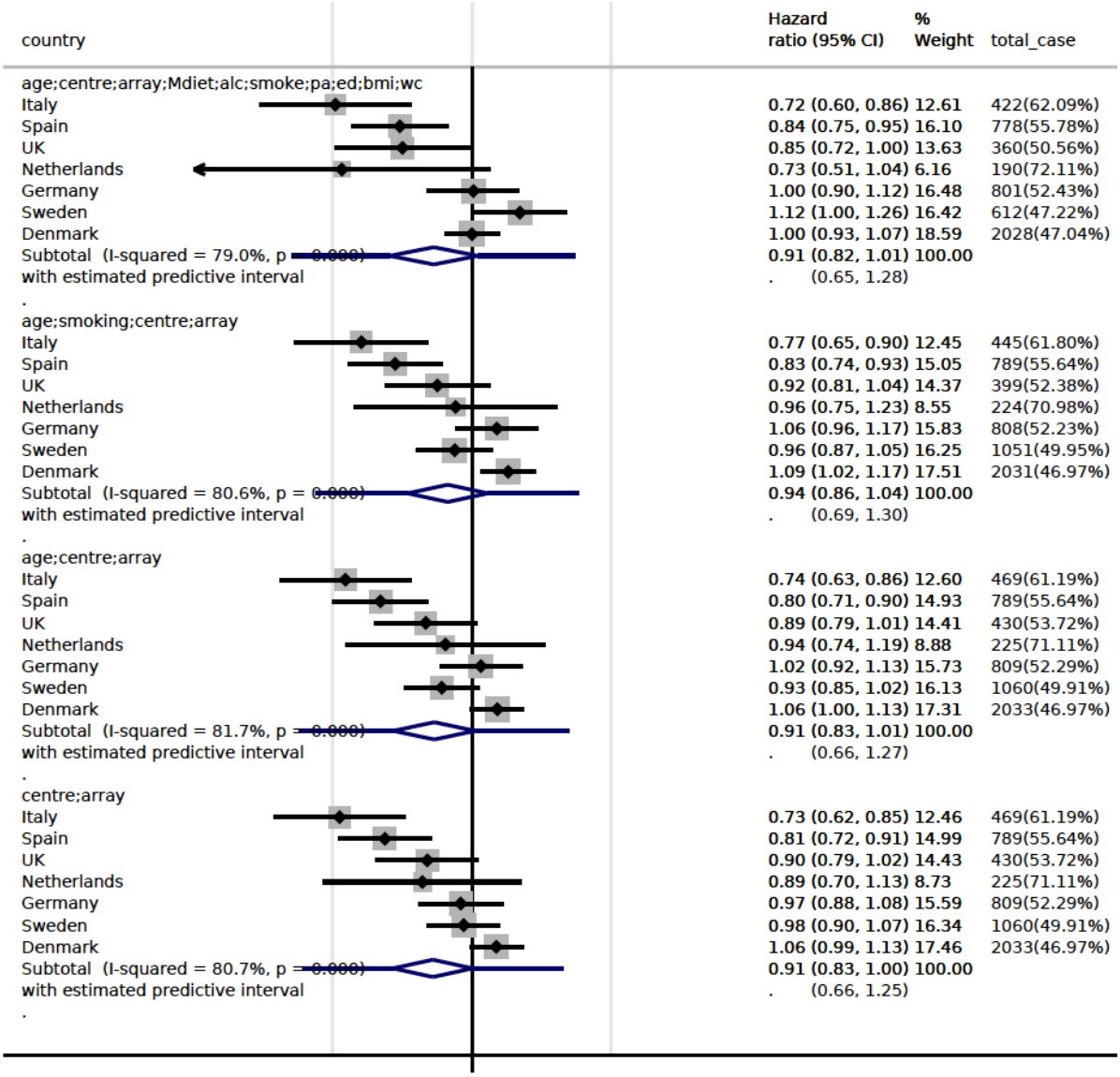
Before data transformation:



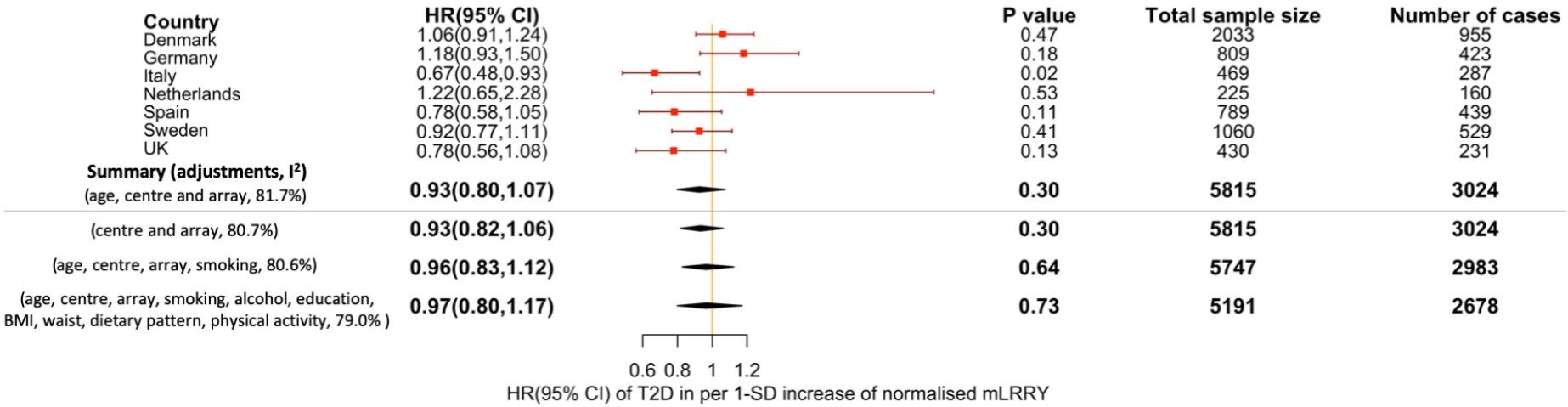
After data transformation:



Supplementary Figure 6: Observational associations between mLRRY and T2D risk. Same models were applied as described in the Figure 5.1, with association estimates shown in each country. *Mdiet*: Mediterranean diet score, *alc*: lifetime alcohol consumption, *pa*: physical activity, ed: educational level, *bmi*: body mass index, *wc*: waist circumference.



Supplementary Figure 7: Observational associations between mLRRY and T2D risk. Same models were applied as described in the Figure 5.1, but with mLRRY as a binary variable based on mLOY indicator (mLRRY>0, i.e. indicating higher, positive values of mLRRY).



Supplementary Tables

Supplementary Table 1: Cohort demographics and LTL measurement data. T/S distributions are given from the primary data prior to z-transformation. Level of statistical significance is denoted by *p<0.01, **p<0.0001. All cohorts showed expected age-associated decline in LTL and higher LTL in women compared to men, except In FINNRISK and NTR_GO2 cohorts, the gender effect was not significant, most likely due to small sample sizes. For the measurement laboratory: 1, Leicester; 2, Helsinki; 3, London; 4, Genetic Laboratory Erasmus MC, Rotterdam; 5, laboratory of Telomere Diagnostics Inc., CA, USA; 6, Cambridge. The inter-run coefficient of variation (CV) is given for LTL measurements performed on triplicates of the same samples.

Cohort	Nationality	Cohort Type	N	Age distribution Mean +/- SD (Range)	Sex distribution % Male	T/S distribution Mean +/- SD (Range)	T/S change per year	Sex effect	LTL lab	LTL CV (%)
EPIC-InterAct T2D cases	Europe	T2D case-cohort	8,499	55.6±7.7 (29-77)	49.16	0.88±0.05(0.67-1.33)	-0.01**	0.15**	6	4.8
EPIC-InterAct subcohort	Europe	T2D case-cohort	12,242	52.5±9.2 (21-79)	37.50	0.88±0.06(0.61-1.36)	-0.016**	0.10**	6	4.8
EPIC-CVD CHD cases	Europe	CVD case-cohort	7,713	59.1±8.4 (21-81)	60.98	0.90±0.09 (0.63-2.11)	-0.001**	0.01**	6	9.8
EPIC-CVD CRBV cases	Europe	CVD case-cohort	3,450	59.2±8.4 (22-84)	50.96	0.88±0.06 (0.65-1.54)	-0.001**	0.01**	6	6.9
EPIC-CVD controls	Europe	CVD case-cohort	752	51.7±12.2 (25-78)	38.70	0.98±0.16 (0.71-2.05)	-0.002**	0.02	6	16.3
BHF-FHS	UK	CAD	1,487	60.8±7.9 (36-82)	80.10	1.35±0.22 (0.69 - 2.13)	-0.006**	0.033*	1	3.5
EGCUT_370	Estonia	Population	2,354	39.87±16.3 (18-91)	47.40	1.85±0.33 (0.89-3.52)	-0.026**	0.246**	1	3.7
EGCUT_OMNI	Estonia	Population	2,234	51.93±20.36	45.80	1.73±0.31 (0.87-3.85)	-0.019**	0.172**	1	3.7
ERF	Netherlands	Population Family based	2,836	49.76±14.87	44.50	1.78 ±0.36	-0.008**	0.068**	1	3.5
FINRISK	Finland	Population	502	51.9±13.8 (25-74)	46.20	1.18± 0.22 (0.69 - 1.93)	-0.0058**	0.035	2	7.7
FTC/NAG-FIN	Finland	Twin, Smokers	831	54.9±4.5 (42-66)	60.40	0.94±0.17 (0.53 - 1.65)	-0.00052	0.041*	2	8.2
GRAPHIC	UK	Population	1,011	52.8±4.40 (40-61)	50.05	1.51±0.23 (0.50 - 2.35)	-0.008**	0.052**	1	3.6
GENMETS cases	Finland	Metabolic syndrome cases	807	51.1±11.1 (30-75)	49.30	1.07±0.20 (0.54 - 1.71)	-0.0047**	0.015	2	na
GENMETS controls	Finland	Metabolic syndrome controls	1,205	51.0±11.0 (30-75)	48.40	1.07±0.20 (0.51 - 1.77)	-0.0037**	0.033*	2	na
HBCS	Finland	Population	1,699	61.5±2.9 (56-69)	42.70	1.39±0.30 (0.32 - 2.59)	-0.013**	0.044*	2	24.8
KORA F3	Germany	Population	3,047	57.1±12.9 (34-85)	48.80	1.72±0.29 (0.92-3.23)	-0.008**	0.053**	1	3.6
KORA F4	Germany	Population	2,907	56.2±13.3 (31-82)	48.30	1.85±0.33 (0.53-3.29)	-0.010**	0.096**	1	3.1
LLS	Netherlands	Population	2,320	59.19± 6.8 (30-80)	45.20	1.46±0.27 (0.74-2.43)	-0.008**	0.046**	1	2.7

Extra rows are shown on the next page

Cohort	Nationality	Cohort Type	N	Age distribution Mean +/- SD (Range)	Sex distribution % Male	T/S distribution Mean +/- SD (Range)	T/S change per year	Sex effect	LTL lab	LTL CV (%)
NESDA	Netherlands	Psychiatric (depression/anxiety) cohort	2,190	42.5±12.9 (18-65)	33.80	1.10±0.29 (0.38-2.33)	-0.006**	0.062**	5	4.6
NFBC1966	Finland	Population, Birth cohort	5,146	31±0	48.20	1.22±0.48 (0.28-4.88)	N/A	0.059**	3	6.2
NTR	Netherlands	Population, Twin	4,977	42.3±15.6 (12-90)	37.60	2.67±0.49 (0.96 - 4.61)	-0.014**	0.088**	1	3.7
QIMR	Australia	Population, Twin	2,438	24.3±14.9 (6-73)	48.77	3.49±0.61 (1.47-5.72)	-0.017**	0.086*	1	3.9
RSI	Netherlands	Population	1800	73.4±8.3 (55-106)	42.00	0.95±0.19 (0.31-1.79)	-0.006**	0.046**	4	4.5
RSIII	Netherlands	Population	372	62.2±8.9 (48-87)	42.30	0.99±0.14 (0.66-1.60)	-0.004**	0.014	4	4.5
TWINGENE	Sweden	Population, Twin	295	71.8±5.9 (55-91)	0.00	1.43±0.25 (0.96–2.26)	-0.011**	N/A	1	2.9
TWINSUK	UK	Population, Twin	4,899	51.0±13.4 (16-99)	9.00	3.71±0.68 (0.68 – 11.40)	-0.016**	-0.008†	1	3.3
UKBS	UK	Population	1,422	43.4±12.4 (17-69)	48.40	1.80±0.50 (0.80 - 3.01)	-0.009**	0.035*	1	3.5

Supplementary Table 2: Details of genotyping platforms and analysis methods used by each study.

Study	Genotyping Platform	Genotype calling algorithm	Genotyped SNPs	Imputation algorithm	Total SNPs (after QC)	Analysis program	Study-specific covariates
EPIC-InterAct	Human CoreExome,	GenomeStudio	550,601	IMPUTE2	13,200,213	SNPtest	batch, centre, top 4 PCs
	Illumina-660W-Quad		652,061		12,133,512		
EPIC-CVD CHD	Human CoreExome,	GenomeStudio	535,701	IMPUTE2	10,965,134	SNPtest	batch, centre, top 4 PCs
	HumanOmniExpress		816,729		9,350,108		
BHF-FHS	Affymetrix 500K	CHIAMO	470,454	IMPUTE2	13,620,894	SNPtest	-
EGCUT_370	Illumina HumanCNV370 HumanOmniExpress	GenomeStudio	306,817	IMPUTE2	30,071,938	SNPtest	-
EGCUT_OMNI	Illumina HumanCNV370 HumanOmniExpress	GenomeStudio	609,578	IMPUTE2	30,071,915	SNPtest	-
ERF	Illumina6K,	Beadstudio	650,197	Mach/Minimac	8,552,982	SNPtest	Family Structure
	Illumina 318K, Illumina370K,						
	Affymetrix 250K						
FINRISK	Illumina 610 Quad	Illuminus	554,988	IMPUTE2	12,248,535	SNPtest	-
FTC/NAG-FIN	Illumina HumanHap670K	Illuminus	549,060	IMPUTE2	13,142,398	SNPtest	-
Genmets Case	Illumina 610 Quad	Illuminus	555,388	IMPUTE2	12,867,930	SNPtest	-
Genmets Control	Illumina 610 Quad	Illuminus	555,388	IMPUTE2	13,399,633	SNPtest	-
GRAPHIC	HumanOmniExpress-12v1	Illumina	648,651	IMPUTE2	13,293,341	SNPtest	-
HBCS	Illumina HumanHap670K	Illuminus	546,814	IMPUTE2	13,806,578	SNPtest	-
KORA F3	Illumina Omni 2.5	GenomeStudio	600,641	IMPUTE v2.3.0	19,805,480	SNPtest	-
	Illumina Omni Express						
KORA F4	Affymetrix Axiom	Affymetrix software	523,260	IMPUTE v2.3.0	20,283,581	SNPtest	-
LLS	Illumina 660w-quad / IlluminaOmniExpress	GenomeStudio	298,538	IMPUTE2	13,382,214	QT-assoc	Family Structure
NESDA	Perlegen-Affymetrix 5.0 Affymetrix 6.0	Perlegen , Afymterix softwares	733,592	Mach/Minimac	8,957,775	SNPtest	chip, top 3 PCs
NFBC1966	Illumina HumanCNV370DUO	Beadstudio	339,629	IMPUTE2	12,253,310	SNPtest	Sex, Age, Plate, Top 3 PCs

NTR	Illumina 370 Affy-Perlegen 5.0, Affy 6.0, Illumina 370, 660, Omni Express 1M	Affymetrix Proprietary Birdseed 1 and 2	289598-1139672	IMPUTE 2.1.2	8,359,471	plink	batch, genotyping chips, PC1, family structure
QIMR	Illumina HumanHap610K	Beadstudio	529,721	Mach 1.0.16, 1.0.18/Minimac 1	10,698,900	merlin-offline	-
RSI	Illumina HumanHap550K	Beadstudio	502,668	Mach/Minimac	11,742,045	mach2qtl	-
RSIII	Illumina HumanHap610Q	Beadstudio	517,658	Mach/Minimac	11,618,162	mach2qtl	-
TWINGENE	Illumina HumanHap300	Beadstudio	307,609	IMPUTE v2.3.0	11,658,532	SNPTEST v2.4.1	-
TWINSUK	Illumina HumanHap300	Illuminus	303,940	IMPUTE2	14,573,410	SNPtest	Family Structure
	Illumina HumanHap610Q		553,487				
	Illumina 1M-Duo		874,733				
UKBS	Affymetrix 500K	CHIAMO	470,398	IMPUTE2	13,663,176	SNPtest	-

Supplementary Table 3: LD between sentinel variants for previously reported loci. LD (R^2 and D') were calculated using LDLink (https://ldlink.nci.nih.gov) between sentinel variants identified in this study and those previously reported. These are broken down by ancestry of the populations from reported studies. LD is calculated for both Europeans (CEU) and for the reported ancestries (CHS or BEB) based on 1000 genomes information.

Population	Chr	Gene	New Lead	Previous lead	R2/D' to new	R2/D' to new lead	R2/D' to new
	3	TERC	rs10936600	rs10936599	1.0/1.0		
	4	NAF1	rs4691895	rs7675998	0.97/1.0		
	5	TERT	rs7705526	rs2736100	0.46/1.0		
European	5	TERT	rs2853677	rs2736100	0.41/0.80		
	10	STN1 (OBFC1)	rs9419958	rs9420907	1.0/1.0		
	14	DCAF4	rs2302588	rs2535913	0.05/1.0		
	19	ZNF208	rs8105767	rs8105767	-		
	20	RTEL1	rs75691080	rs755017	0.01/1.0		
	20	RTEL1	rs34978822	rs755017	0.004/0.22		
	20	RTEL1	rs73624724	rs755017	0.89/1.0		
Singaporean Chinese	1	PARP1	rs3219104	rs3219104	-	-	
	7	POT1	rs59294613	rs7776744	0.23/0.87	0.43/1.0	
	11	ATM	rs228595	rs227080	0.42/0.91	0.83/0.92	
	16	MPHOSPH6	rs7194734	rs2967374	0.95/0.97	1.0/1.0	
	18	TYMS	rs2124616	rs1001761	0.27/1.0	0.002/1.0	
	20	RTEL1	rs75691080	rs41309367	0.03/0.80	0.002/1.0	
	20	RTEL1	rs34978822	rs41309367	0.03/1.0	NA	
	20	RTEL1	rs73624724	rs41309367	0.25/0.85	0.01/0.21	
South Asian	20	RTEL1	rs75691080	rs2297439	0.50/0.80		0.04/0.39
	20	RTEL1	rs34978822	rs2297439	0.001/1.0		NA
	20	RTEL1	rs73624724	rs2297439	0.002/0.44		0.06/0.87

Supplementary Table 4: Independent variants associated with LTL at FDR<0.05. Columns indicate (Chr) chromosome ; SNP; (bp) physical position (hg19); (freq) frequency of the effect allele in the original GWAS data; (refA) the effect allele; (b) effect size, (se) standard error and (p) *p*-value from single variant based GWAS meta-analysis; (n) estimated effective sample size; (freq_geno) frequency of the effect allele in the reference sample; (bJ),(bJ_se), (pJ) effect size, standard error and *p*-value from joint models; and (LD_r) between the variant and the locus sentinel variant.

Chr	SNP	bp	refA	freq	b	se	p	n	freq_geno	bJ	bJ_se	pJ	LD_r
3	rs10936600	169514585	T	0.243	-0.0858	0.0057	6.42E-51	80402	0.243	-0.0858	0.0057	8.79E-51	0
5	rs7705526	1285974	A	0.328	0.0820	0.0058	4.82E-45	64656.3	0.324	0.0662	0.0063	3.55E-26	-0.36568
5	rs2853677	1287194	A	0.592	-0.0638	0.0055	3.12E-31	66348.1	0.576	-0.0413	0.0059	2.41E-12	0
4	rs4691895	164048199	C	0.783	0.0577	0.0061	1.47E-21	77751.1	0.775	0.0577	0.0061	1.55E-21	0
10	rs9419958	105675946	C	0.862	-0.0636	0.0071	4.77E-19	79673.7	0.869	-0.0636	0.0071	4.96E-19	0
20	rs75691080	62269750	T	0.091	-0.0671	0.0089	5.75E-14	73299.7	0.085	-0.0636	0.0090	1.44E-12	-0.04694
7	rs59294613	124554267	A	0.293	-0.0407	0.0055	1.12E-13	77807.4	0.290	-0.0407	0.0055	1.14E-13	0
19	rs8105767	22215441	G	0.289	0.0392	0.0054	5.21E-13	80103	0.289	0.0392	0.0054	5.30E-13	0
20	rs73624724	62436398	C	0.129	0.0507	0.0074	6.08E-12	79451.3	0.137	0.0390	0.0075	2.07E-07	0
1	rs3219104	226562621	C	0.830	0.0417	0.0064	9.31E-11	82701.8	0.847	0.0417	0.0064	9.41E-11	0
20	rs932827	62380527	T	0.238	-0.0374	0.0060	3.28E-10	75271.4	0.229	-0.0308	0.0061	4.31E-07	-0.18363
6	rs2736176	31587561	C	0.313	0.0345	0.0055	3.41E-10	74733.4	0.284	0.0322	0.0055	5.18E-09	0
16	rs3785074	69406986	G	0.263	0.0351	0.0056	4.50E-10	78946.7	0.284	0.0350	0.0056	5.01E-10	-0.00293
16	rs7194734	82199980	T	0.782	-0.0369	0.0060	6.72E-10	79221.3	0.770	-0.0372	0.0060	5.39E-10	0
20	rs34978822	62291599	G	0.015	-0.1397	0.0227	7.04E-10	64578.6	0.021	-0.1486	0.0228	7.04E-11	-0.07009
6	rs34991172	25480328	G	0.068	-0.0608	0.0105	6.03E-09	69563.3	0.082	-0.0560	0.0105	9.24E-08	-0.08086
11	rs228595	108105593	A	0.417	-0.0285	0.0050	1.39E-08	79131.2	0.411	-0.0285	0.0050	1.40E-08	0
14	rs2302588	73404752	C	0.100	0.0476	0.0084	1.64E-08	75515	0.103	0.0482	0.0084	1.07E-08	-0.02093
4	rs13137667	71774347	C	0.959	0.0765	0.0137	2.37E-08	65743.6	0.974	0.0765	0.0137	2.39E-08	0
3	rs55749605	101232093	A	0.579	-0.0373	0.0067	2.38E-08	44477.5	0.622	-0.0373	0.0067	2.41E-08	0
16	rs62053580	74680074	G	0.169	-0.0389	0.0071	3.96E-08	68784.9	0.169	-0.0390	0.0071	3.48E-08	-0.00711
2	rs754017156	54482703	D	0.165	0.0471	0.0088	7.52E-08	45835	0.146	0.0471	0.0088	7.59E-08	0
15	rs12909131	50387678	T	0.231	-0.0308	0.0058	1.15E-07	80706.5	0.239	-0.0310	0.0058	9.60E-08	-0.01047
20	rs1744757	35734863	T	0.851	0.0359	0.0068	1.38E-07	82222.6	0.852	0.0359	0.0068	1.38E-07	0
18	rs2124616	661917	A	0.140	-0.0374	0.0072	1.72E-07	78571.2	0.153	-0.0374	0.0072	1.73E-07	0
3	rs2613954	112847045	T	0.886	-0.0381	0.0078	1.10E-06	78132.7	0.878	-0.0381	0.0078	1.11E-06	0

1	rs12065882	114078755	G	0.208	0.0298	0.0062	1.36E-06	77170.9	0.207	0.0298	0.0062	1.37E-06	0
10	rs2386642	5702259	A	0.673	-0.0256	0.0053	1.44E-06	78324.5	0.665	-0.0256	0.0053	1.44E-06	0
2	rs56810761	210663697	T	0.270	0.0275	0.0057	1.45E-06	75729.8	0.268	0.0275	0.0057	1.45E-06	0
5	rs62365174	78925743	G	0.088	-0.0544	0.0113	1.50E-06	47138.2	0.093	-0.0544	0.0113	1.51E-06	0
12	rs112655343	14430807	T	0.102	0.0425	0.0090	2.22E-06	65703.2	0.110	0.0425	0.0090	2.23E-06	0
15	rs55710439	65229816	T	0.014	0.1050	0.0223	2.65E-06	69379.6	0.012	0.1050	0.0224	2.66E-06	0
16	rs11640926	1249877	G	0.139	0.0557	0.0119	2.93E-06	28512.8	0.140	0.0557	0.0119	2.95E-06	0
4	rs60160057	151000830	A	0.211	-0.0287	0.0062	3.15E-06	76458.6	0.219	-0.0287	0.0062	3.16E-06	0
14	rs117536281	105494403	G	0.034	0.0850	0.0183	3.31E-06	43901.3	0.035	0.0850	0.0183	3.33E-06	0
22	rs7510583	44698803	G	0.290	0.0347	0.0075	3.38E-06	42136.9	0.280	0.0347	0.0075	3.40E-06	0
14	rs59192843	74514120	G	0.059	0.0655	0.0141	3.52E-06	43632	0.043	0.0668	0.0141	2.28E-06	0
8	rs57415150	2882469	A	0.042	-0.0584	0.0126	3.68E-06	76209.6	0.040	-0.0584	0.0126	3.69E-06	0
20	rs6038821	7402809	T	0.038	0.0596	0.0129	3.98E-06	78795.1	0.025	0.0593	0.0129	4.49E-06	0
17	rs144204502	76183233	T	0.014	-0.0896	0.0196	4.92E-06	90239	0.012	-0.0896	0.0196	4.94E-06	0
20	rs6107615	5310273	C	0.422	-0.0228	0.0050	5.30E-06	79235.8	0.422	-0.0227	0.0050	5.98E-06	-0.00545
15	rs9972513	38930961	T	0.281	0.0247	0.0055	5.75E-06	80585.1	0.278	0.0247	0.0055	5.76E-06	0
11	rs117037102	93404608	T	0.018	0.0979	0.0218	6.81E-06	58251	0.021	0.0979	0.0218	6.83E-06	0
21	rs7276273	45994841	C	0.007	-0.1502	0.0334	6.90E-06	58815.8	0.010	-0.1502	0.0334	6.92E-06	0
19	rs11665818	39768216	A	0.195	0.0278	0.0062	7.04E-06	80994.7	0.188	0.0278	0.0062	7.06E-06	0
14	rs3213718	90869913	T	0.583	0.0224	0.0050	7.22E-06	79728.4	0.602	0.0224	0.0050	7.24E-06	0
5	rs112347796	138964816	D	0.049	0.0691	0.0154	7.29E-06	43935.8	0.054	0.0691	0.0154	7.32E-06	0
19	rs143276018	3939249	C	0.018	-0.1015	0.0229	9.02E-06	51875.2	0.015	-0.1015	0.0229	9.06E-06	0
8	rs201375979	100917632	D	0.317	0.0332	0.0075	9.11E-06	39878.3	0.358	0.0332	0.0075	9.15E-06	0
12	rs7311314	54592103	A	0.317	0.0240	0.0054	9.50E-06	75916	0.309	0.0240	0.0054	9.52E-06	0
1	rs35675808	167399643	G	0.028	0.0736	0.0166	9.54E-06	64171.8	0.022	0.0736	0.0166	9.57E-06	0
15	rs117610974	55105443	G	0.009	-0.1540	0.0350	1.05E-05	42498.8	0.010	-0.1555	0.0350	8.74E-06	0

Supplementary Table 5: Comparison of all loci at FDR<0.05 to that reported in the Singaporean Chinese Health Study (SCHS). Data is sorted by original *p*-value, pJ indicates *p*-value from conditional (GCTA) analyses. Minor allele frequencies (MAF) are given from 1000 genomes populations for information. Variants with MAF<0.01 were excluded in the SCHS study so not available. Many of our variants were monoallelic in the SCHS and denoted by" -". Variants that were only genotyped in our study but not in the SCHS dataset or 1000 genomes reference panel, were denoted by "NA".

									SCHS data								Allele frequencies	
Chr	SNP	bp	ref Allele	Closest gene (prioritised)	beta	se	p	pJ	rsid_SCHS	If proxy used		reported allele	p	beta	se	p_het	MAF CEU	MAF CSH
										R^2 CEU	R^2 CSH							
3	rs10936600	169514585	T	LRRC34	-0.09	0.01	6.42E-51	8.79E-51	rs10936600			T	1.85E-38	-0.12	0.01	0.57	0.26	0.47
5	rs7705526	1285974	A	TERT	0.08	0.01	4.82E-45	3.55E-26	rs7705526			A	2.61E-38	0.12	0.01	0.06	0.32	0.34
5	rs2853677	1287194	A	TERT	-0.06	0.01	3.12E-31	2.41E-12	rs2853677			A	2.73E-29	-0.10	0.01	0.60	0.40	0.33
4	rs4691895	164048199	C	NAF1	0.06	0.01	1.47E-21	1.55E-21	rs4691895			C	1.36E-08	0.06	0.01	0.70	0.22	0.22
10	rs9419958	105675946	C	STN1 (OBFC1)	-0.06	0.01	4.77E-19	4.96E-19	rs9419958			C	0.247663	-0.04	0.04	0.66	0.14	0.02
20	rs75691080	62269750	T	STMN3	-0.07	0.01	5.75E-14	1.44E-12	-								0.07	0.005
7	rs59294613	124554267	A	POT1	-0.04	0.01	1.12E-13	1.14E-13	rs59294613			A	0.000391	-0.03	0.01	0.97	0.26	0.36
19	rs8105767	22215441	G	ZNF257	0.04	0.01	5.21E-13	5.30E-13	rs8105767			G	0.000221	0.04	0.01	0.20	0.29	0.3
20	rs73624724	62436398	C	ZBTB46	0.05	0.01	6.08E-12	2.07E-07	rs73624724			C	0.840162	0.00	0.01	0.03	0.16	0.48
1	rs3219104	226562621	C	PARP1	0.04	0.01	9.31E-11	9.41E-11	rs3219104			C	2.43E-16	0.07	0.01	0.32	0.14	0.44
20	rs932827	62380527	T	ZBTB46	-0.04	0.01	3.28E-10	4.31E-07	rs932827			T	0.001667	-0.05	0.02	0.30	0.24	0.07
6	rs2736176	31587561	C	PRRC2A (CSNK2B, BAG6)	0.03	0.01	3.41E-10	5.18E-09	rs2736176			C	0.034688	0.02	0.01	0.85	0.30	0.38
16	rs3785074	69406986	G	TERF2	0.04	0.01	4.50E-10	5.01E-10	rs3785074			G	5.78E-05	0.06	0.02	0.20	0.30	0.12
16	rs7194734	82199980	T	MPHOSPH6	-0.04	0.01	6.72E-10	5.39E-10	rs7194734			T	5.84E-06	-0.06	0.01	0.62	0.24	0.19
20	rs34978822	62291599	G	RTEL1	-0.14	0.02	7.04E-10	7.04E-11	-								0.02	-
6	rs34991172	25480328	G	CARMIL1	-0.06	0.01	6.03E-09	9.24E-08	-								0.09	-
11	rs228595	108105593	A	ATM	-0.03	0.01	1.39E-08	1.40E-08	rs228595			A	1.11E-07	-0.05	0.01	0.54	0.37	0.44
14	rs2302588	73404752	C	DCAF4	0.05	0.01	1.64E-08	1.07E-08	rs2302588			C	0.000127	0.04	0.01	0.38	0.11	0.22
4	rs13137667	71774347	C	MOB1B (DCK)	0.08	0.01	2.37E-08	2.39E-08	rs13137667			C	0.027597	0.05	0.02	0.35	0.05	0.04
3	rs55749605	101232093	A	SENP7	-0.04	0.01	2.38E-08	2.41E-08	rs55749605			A	0.134509	-0.01	0.01	0.14	0.37	0.34
16	rs62053580	74680074	G	RFWD3	-0.04	0.01	3.96E-08	3.48E-08	rs62053580			G	0.0224	-0.02	0.01	0.19	0.14	0.29

Extra rows are shown on the next page

									SCHS data								Allele frequencies	
Chr	SNP	bp	ref Allele	Closest gene (prioritised)	beta	se	p	pJ	rsid_SCHS	If proxy used		reported allele	p	beta	se	p_het	MAF CEU	MAF CSH
										R^2 CEU	R^2 CSH							
2	rs754017156	54482703	D	ACYP2	0.05	0.01	7.52E-08	7.59E-08	rs1872329	1	0.91	A	0.001684	0.04	0.01	0.74	0.16	0.19
15	rs12909131	50387678	T	ATP8B4	-0.03	0.01	1.15E-07	9.60E-08	rs12909131			T	0.028708	-0.02	0.01	0.47	0.21	0.24
20	rs1744757	35734863	T	MROH8 (SAMHD1)	0.04	0.01	1.38E-07	1.38E-07	rs1744757			T	0.002198	0.03	0.01	0.06	0.17	0.49
18	rs2124616	661917	A	TYMS	-0.04	0.01	1.72E-07	1.73E-07	-								0.16	0.005
3	rs2613954	112847045	T	RP11-572M11.4	-0.04	0.01	1.10E-06	1.11E-06									0.12	0.0095
1	rs12065882	114078755	G	MAGI3	0.03	0.01	1.36E-06	1.37E-06	rs12065882			G	0.691999	0.01	0.03	0.25	0.17	0.02
10	rs2386642	5702259	A	ASB13	-0.03	0.01	1.44E-06	1.44E-06	rs2386642			A	0.768391	0.00	0.01	0.15	0.32	0.28
2	rs56810761	210663697	T	UNC80	0.03	0.01	1.45E-06	1.45E-06	rs56810761			T	0.003973	0.04	0.01	0.27	0.25	0.16
5	rs62365174	78925743	G	PAPD4	-0.05	0.01	1.50E-06	1.51E-06	rs62365174			G	0.047553	-0.02	0.01	0.56	0.10	0.14
12	rs112655343	14430807	T	ATF7IP	0.04	0.01	2.22E-06	2.23E-06	-								0.11	-
15	rs55710439	65229816	T	ANKDD1A	0.10	0.02	2.65E-06	2.66E-06	-								0.01	-
16	rs11640926	1249877	G	CACNA1H	0.06	0.01	2.93E-06	2.95E-06	-								0.12	0.13 (not in SCHC dataset)

Extra rows are shown on the next page

									SCHS data								Allele frequencies	
Chr	SNP	bp	ref Allele	Closest gene (prioritised)	beta	se	p	pJ	rsid_SCHS	If proxy used		reported allele	p	beta	se	p_het	MAF CEU	MAF CSH
										R^2 CEU	R^2 CSH							
4	rs60160057	151000830	A	DCLK2	-0.03	0.01	3.15E-06	3.16E-06	rs60160057			A	0.554438	-0.01	0.02	0.91	0.27	0.11
14	rs117536281	105494403	G	CDCA4	0.08	0.02	3.31E-06	3.33E-06	-								0.04	-
22	rs7510583	44698803	G	KIAA1644	0.03	0.01	3.38E-06	3.40E-06	-								NA	NA
14	rs59192843	74514120	G	CCDC176	0.07	0.01	3.52E-06	2.28E-06	rs59192843			G	0.423169	-0.01	0.02	0.69	0.06	0.18
8	rs57415150	2882469	A	CSMD1	-0.06	0.01	3.68E-06	3.69E-06	rs57415150			A	0.040218	-0.04	0.02	0.30	0.06	0.09
20	rs6038821	7402809	T	LINC01706	0.06	0.01	3.98E-06	4.49E-06	rs6038821			T	0.878709	0.00	0.01	0.07	0.04	0.30
17	rs144204502	76183233	T	TK1	-0.09	0.02	4.92E-06	4.94E-06	-								0.01	-
20	rs6107615	5310273	C	PROKR2	-0.02	0.01	5.30E-06	5.98E-06	rs6107615			C	0.712051	0.00	0.01	0.80	0.45	0.26
15	rs9972513	38930961	T	RP11-275I4.2	0.02	0.01	5.75E-06	5.76E-06	-								NA	NA
11	rs117037102	93404608	T	CEP295	0.10	0.02	6.81E-06	6.83E-06	-								0.005	-
21	rs7276273	45994841	C	KRTAP10-4	-0.15	0.03	6.90E-06	6.92E-06	-								0.03	-
19	rs11665818	39768216	A	IFNL2	0.03	0.01	7.04E-06	7.06E-06	rs11665818			A	0.392	-0.02	0.03	0.62	0.19	0.04
14	rs3213718	90869913	T	CALM1	0.02	0.00	7.22E-06	7.24E-06	rs3213718			T	0.166871	-0.02	0.01	0.81	0.39	0.2
5	rs112347796	138964816	D	UBE2D2	0.07	0.02	7.29E-06	7.32E-06	-								NA	NA
19	rs143276018	3939249	C	NMRK2	-0.10	0.02	9.02E-06	9.06E-06	-								0.03	-
8	rs201375979	100917632	D	COX6C	0.03	0.01	9.11E-06	9.15E-06	rs10098852	1	1	G	0.709197	0.00	0.01	0.52	0.39	0.46
12	rs7311314	54592103	A	SMUG1	0.02	0.01	9.50E-06	9.52E-06	rs7311314			A	0.331608	-0.01	0.01	0.65	0.23	0.4
1	rs35675808	167399643	G	CD247	0.07	0.02	9.54E-06	9.57E-06	-								0.02	-
15	rs117610974	55105443	G	UNC13C	-0.15	0.03	1.05E-05	8.74E-06	-								0.03	-

Supplementary Table 6: Functional prediction of nonsynonymous variants. Coding variants were identified within each locus with $r^2 \geqslant 0.8$ to the locus lead SNP. Functional prediction of the amino acid changes was carried out using PolyPhen, SIFT and CADD prediction tools. CADD scores above 20 are considered to be within the 1% most deleterious mutations. PD: probably damaging; B: benign; U: unknown; T: tolerance; D: damaging.

Chr	Lead SNP	SNP	r2	Variant	Gene	Transcript	AA	wild	mutant	POLYPHEN		SIFT			CADD
							Position	AA	AA	*Score*	*Prediction*	*Score*	*Prediction*	*Confidence*	
1	rs3219104	rs1136410	1	A|G	PARP1	ENST00000366794	762	V	A	0.827	PD	0.24	T	HIGH	28.1
		rs1805415	1	T|G		ENST00000366794	352	K	N	0.059	B	0.47	T	HIGH	10.25
3	rs55749605	rs2433031	1	T|A	SENP7	ENST00000394095	612	Q	H	0.678	PD	0.55	T	HIGH	23.3
						ENST00000394091	448	Q	H	0.413	PD	0.54	T	HIGH	23.3
						ENST00000394094	547	Q	H	0.948	PD	0.54	T	HIGH	23.3
						ENST00000314261	546	Q	H	0.678	PD	0.55	T	HIGH	23.3
						ENST00000348610	579	Q	H	0.678	PD	0.58	T	HIGH	23.3
						ENST00000366089	14	Q	H	0.433	PD	0.25	T	HIGH	23.3
						ENST00000358203	448	Q	H	0.413	PD	0.54	T	HIGH	23.3
3	rs10936600	rs10936600	-	A|T	LRRC34	ENST00000446859	286	L	I	0.863	PD	0.69	T	HIGH	14.74
						ENST00000522830	225	L	I	0.93	PD	0.53	T	HIGH	14.74
						ENST00000522526	254	L	I	0.863	PD	0.42	T	HIGH	14.74
						ENST00000528597	35	L	I	0.958	PD	0.06	T	HIGH	14.74
						ENST00000316515	241	L	I	0.93	PD	0.44	T	HIGH	14.74
		rs6793295	0.93	T|C	LRRC34	ENST00000446859	249	S	G	0	B	0.51	T	HIGH	11.05
						ENST00000522830	188	S	G	0	B	0.52	T	HIGH	11.05
4	rs4691895	rs4691895	-	G|C	NAF1	ENST00000422287	368	L	V	0	B	0.66	T	HIGH	0.505
		rs4691896	1	T|C		ENST00000422287	162	I	V	0	B	0.33	T	HIGH	3.449
						ENST00000274054	162	I	V	0	B	0.3	T	HIGH	
11	rs117037102	rs117405490	1	C|G	CEP295	ENST00000325212	783	P	A	0.907	PD	0.24	T	HIGH	11.9
						ENST00000411936	783	P	A	0.907	PD	0.27	T	HIGH	11.9

14	rs2302588	rs2302588	-	G|C	DCAF4	ENST00000358377	22	W	C	0.993	PD	0.01	D	LOW	14.81
						ENST00000353777	22	W	C	0	U	0	D	LOW	14.81
						ENST00000509320	22	W	C	0	U	0.04	D	LOW	14.81
						ENST00000509153	22	W	C	0.993	PD	0.01	D	LOW	14.81
		rs3815460	1	C|G		ENST00000358377	345	S	C	0.995	PD	0.03	D	HIGH	28.5
						ENST00000353777	175	S	C	0.994	PD	0.05	D	HIGH	28.5
						ENST00000394234	245	S	C	0.995	PD	0	D	HIGH	28.5
						ENST00000509153	285	S	C	0.998	PD	0.01	D	HIGH	28.5
		rs2286838	0.9	G|C	ZFYVE1	ENST00000318876	408	S	R	0.788	PD	0	D	HIGH	3.451
16	rs7194734	rs2303262	0.95	C|T	MPHOSPH6	ENST00000258169	8	R	K	0	B	1	T	HIGH	19.52
20	rs34978822	rs35640778	1	G|A	RTEL1	ENST00000370018	684	R	Q	0.008	B	0.69	T	HIGH	19.92
						ENST00000508582	708	R	Q	0.008	B	0.71	T	HIGH	19.92
						ENST00000360203	684	R	Q	0.008	B	0.76	T	HIGH	19.92
						ENST00000425905	77	R	Q	0.008	B				19.92
						ENST00000318100	684	R	Q	0.02	B	0.63	T	HIGH	19.92
						ENST00000492259	712	R	Q	0.08	B				19.92
						ENST00000482936	684	R	Q	0.02	B				19.92
20	rs73624724	rs2281929	0.89	T|C	ZBTB46	ENST00000245663	11	T	A	0	B	0.36	T	HIGH	11.68
						ENST00000302995	11	T	A	0	B	0.36	T	HIGH	11.68
						ENST00000395104	11	T	A	0	B	0.36	T	HIGH	11.68

Supplementary Table 7: Integration of eQTLs using S-PrediXcan and co-localisation analyses. Genes are identified by Ensembl IDs and gene names are derived from the UCSC Human Genome database. Genes were allocated to overlapping LTL loci where possible, with sentinel SNPs of the corresponding loci shown. Detailed column specifications were given in software websites (section 2.2.6.2).

Sentinel SNP and tissue-specific gene expression							Co-localisation				
SNP	CHR	Gene_start	Gene_end	Gene name	Gene nsnps	Tissue	H0_abf	H1_abf	H2_abf	H3_abf	H4_abf
rs12065882	1	114437370	114447762	AP4B1	5041	Whole_Blood	0.00	0.00	0.01	0.03	0.97
rs12065882	1	114437370	114447762	AP4B1	4998	Uterus	0.00	0.02	0.01	0.05	0.91
rs12065882	1	114437370	114447762	AP4B1	5041	Heart_Left_Ventricle	0.00	0.00	0.03	0.15	0.82
rs12065882	1	114437370	114447762	AP4B1	5041	Muscle_Skeletal	0.00	0.00	0.03	0.15	0.81
rs12065882	1	114399257	114443857	AP4B1-AS1	5024	Nerve_Tibial	0.01	0.03	0.01	0.06	0.88
rs12065882	1	114399257	114443857	AP4B1-AS1	5024	Skin_Not_Sun_Exposed_Suprapubic	0.01	0.02	0.03	0.11	0.83
rs12065882	1	114399257	114443857	AP4B1-AS1	5024	Thyroid	0.00	0.01	0.03	0.13	0.82
rs12065882	1	114399257	114443857	AP4B1-AS1	5024	Whole_Blood	0.00	0.00	0.02	0.08	0.89
rs12065882	1	114239453	114302111	PHTF1	5069	Muscle_Skeletal	0.00	0.02	0.02	0.08	0.88
rs12065882	1	114356433	114414381	PTPN22	5033	Brain_Cerebellum	0.00	0.00	0.02	0.08	0.90
rs12065882	1	114356433	114414381	PTPN22	5038	Colon_Transverse	0.00	0.00	0.01	0.05	0.94
rs12065882	1	114356433	114414381	PTPN22	5038	Pancreas	0.00	0.00	0.01	0.07	0.92
rs3219104	1	226736501	226796915	C1orf95	6156	Brain_Anterior_cingulate_cortex_BA24	0.00	0.01	0.00	0.04	0.95
rs3219104	1	226548392	226595780	PARP1	6127	Pancreas	0.00	0.00	0.00	0.03	0.97
rs754017156	2	54480315	54483409	TSPYL6	7075	Testis	0.00	0.00	0.00	0.04	0.96
rs56810761	2	210673528	210674304	SNAI1P1	5605	Testis	0.00	0.00	0.03	0.13	0.84
rs55749605	3	101043049	101232085	SENP7	5676	Small_Intestine_Terminal_Ileum	0.00	0.01	0.01	0.15	0.83
rs10936600	3	169490619	169507504	MYNN	5774	Artery_Aorta	0.00	0.22	0.00	0.30	0.48
rs10936600	3	169490619	169507504	MYNN	5774	Testis	0.00	0.00	0.00	0.02	0.98
rs10936600	3	169511216	169530774	LRRC34	5802	Adipose_Subcutaneous	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Adipose_Visceral_Omentum	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5800	Adrenal_Gland	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Artery_Aorta	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Artery_Tibial	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5801	Cells_Transformed_fibroblasts	0.00	0.00	0.00	0.97	0.03
rs10936600	3	169511216	169530774	LRRC34	5801	Colon_Sigmoid	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5801	Esophagus_Gastroesophageal_Junction	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Esophagus_Mucosa	0.00	0.38	0.00	0.59	0.04
rs10936600	3	169511216	169530774	LRRC34	5802	Heart_Atrial_Appendage	0.00	0.50	0.00	0.32	0.18
rs10936600	3	169511216	169530774	LRRC34	5801	Heart_Left_Ventricle	0.00	0.00	0.00	0.96	0.04
rs10936600	3	169511216	169530774	LRRC34	5802	Muscle_Skeletal	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Nerve_Tibial	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169511216	169530774	LRRC34	5802	Thyroid	0.00	0.00	0.00	1.00	0.00
rs10936600	3	169539710	169555563	LRRIQ4	5821	Cells_Transformed_fibroblasts	0.00	0.04	0.00	0.95	0.01
rs10936600	3	169539710	169555563	LRRIQ4	5820	Testis	0.00	0.00	0.00	1.00	0.00
rs60160057	4	150999426	151178609	DCLK2	4857	Spleen	0.00	0.00	0.02	0.09	0.90
rs4691895	4	164031225	164088073	NAF1	7177	Muscle_Skeletal	0.00	0.00	0.00	0.93	0.07
rs4691895	4	164031225	164088073	NAF1	7177	Skin_Sun_Exposed_Lower_leg	0.00	0.27	0.00	0.23	0.50
rs4691895	4	164031225	164088073	NAF1	7177	Thyroid	0.00	0.02	0.00	0.06	0.92
rs4691895	4	164031225	164088073	NAF1	7177	Cells_Transformed_fibroblasts	0.00	0.03	0.00	0.08	0.89
rs4691895	4	164029937	164041117	RP11-563E2.2	7235	Adipose_Subcutaneous	0.00	0.02	0.00	0.06	0.92
rs4691895	4	164029937	164041117	RP11-563E2.2	7235	Nerve_Tibial	0.00	0.00	0.00	0.04	0.96
rs4691895	4	164029937	164041117	RP11-563E2.2	7235	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.04	0.00	0.07	0.89
rs4691895	4	164029937	164041117	RP11-563E2.2	7235	Skin_Sun_Exposed_Lower_leg	0.00	0.01	0.00	0.04	0.95
rs62365174	5	78908243	78982471	PAPD4	6194	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.00	0.02	0.11	0.86
rs62365174	5	78908243	78982471	PAPD4	6194	Breast_Mammary_Tissue	0.00	0.00	0.02	0.11	0.86
rs62365174	5	78908243	78982471	PAPD4	6194	Colon_Transverse	0.00	0.00	0.03	0.15	0.82
rs62365174	5	78908243	78982471	PAPD4	6194	Esophagus_Muscularis	0.00	0.00	0.03	0.13	0.85
rs62365174	5	78908243	78982471	PAPD4	6194	Lung	0.00	0.00	0.02	0.12	0.86
rs62365174	5	78908243	78982471	PAPD4	6194	Nerve_Tibial	0.00	0.00	0.02	0.11	0.86
rs62365174	5	78908243	78982471	PAPD4	6194	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.03	0.12	0.86
rs62365174	5	78908243	78982471	PAPD4	6181	Small_Intestine_Terminal_Ileum	0.00	0.00	0.02	0.12	0.86
rs62365174	5	78908243	78982471	PAPD4	6194	Thyroid	0.00	0.00	0.02	0.12	0.86
rs2736176	6	31082527	31107869	PSORS1C1	17178	Heart_Left_Ventricle	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31110216	31126015	CCHCR1	16891	Colon_Sigmoid	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31110216	31126015	CCHCR1	16891	Nerve_Tibial	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31132119	31148508	POU5F1	16841	Adipose_Visceral_Omentum	0.00	0.00	0.00	0.96	0.04
rs2736176	6	31236526	31239882	HLA-C	17050	Adrenal_Gland	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Artery_Aorta	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17053	Artery_Coronary	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Artery_Tibial	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Breast_Mammary_Tissue	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Colon_Transverse	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Heart_Atrial_Appendage	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17059	Heart_Left_Ventricle	0.00	0.00	0.00	1.00	0.00

Extra columns are shown on the next page

S-PrediXcan (if overlapped with colocalisation)													
Gene name	Tissue	effect size	zscore	pvalue	var_g	pred_perf			BEST_GWAS		n_snps		
						r^2	pval	qval	ID	Z	used	cov	mode
AP4B1	Whole_Blood												
AP4B1	Uterus												
AP4B1	Heart_Left_Ventricle												
AP4B1	Muscle_Skeletal	0.20	5.34	9.55E-08	0.01	0.01	7.98E-02	3.49E-02	rs17464525	-4.60	12	12	12
AP4B1-AS1	Nerve_Tibial												
AP4B1-AS1	Skin_Not_Sun_Exposed_Suprapubic												
AP4B1-AS1	Thyroid												
AP4B1-AS1	Whole_Blood												
PHTF1	Muscle_Skeletal												
PTPN22	Brain_Cerebellum												
PTPN22	Colon_Transverse												
PTPN22	Pancreas												
C1orf95	Brain_Anterior_cingulate_cortex_BA24												
PARP1	Pancreas	0.12	6.10	1.08E-09	0.03	0.06	2.63E-03	3.12E-03	rs2255403	-6.39	4	4	4
TSPYL6	Testis	-0.03	-5.32	1.02E-07	0.47	0.56	1.43E-29	6.56E-28	rs12615793	5.22	30	30	30
SNAI1P1	Testis												
SENP7	Small_Intestine_Terminal_Ileum												
MYNN	Artery_Aorta	-0.09	-6.62	3.69E-11	0.08	0.07	2.58E-04	2.66E-04	rs7621631	-14.97	31	31	31
MYNN	Testis	-0.24	-13.35	1.23E-40	0.04	0.06	1.75E-03	1.46E-03	rs7621631	-14.97	10	10	10
LRRC34	Adipose_Subcutaneous	0.20	13.09	3.58E-39	0.05	0.09	1.28E-07	1.49E-07	rs3821383	-14.15	21	22	22
LRRC34	Adipose_Visceral_Omentum	0.37	12.29	1.09E-34	0.01	0.05	1.32E-03	1.82E-03	rs3821383	-14.15	16	18	18
LRRC34	Adrenal_Gland	0.43	12.31	7.59E-35	0.01	0.08	1.21E-03	2.11E-03	rs9833035	-13.05	7	8	8
LRRC34	Artery_Aorta	0.08	7.30	2.89E-13	0.10	0.12	1.02E-06	1.64E-06	rs6793160	-11.54	32	34	34
LRRC34	Artery_Tibial	0.09	8.85	8.66E-19	0.10	0.08	1.00E-06	1.03E-06	rs6793160	-11.54	43	48	48
LRRC34	Cells_Transformed_fibroblasts	0.14	12.54	4.40E-36	0.10	0.13	4.93E-10	5.88E-10	rs1997392	-14.17	26	30	30
LRRC34	Colon_Sigmoid	0.18	10.98	4.69E-28	0.05	0.19	4.49E-07	2.35E-06	rs10936596	-11.97	16	17	17
LRRC34	Esophagus_Gastroesophageal_Junction	0.15	8.46	2.58E-17	0.04	0.13	3.43E-05	1.25E-04	rs9878797	-10.62	10	11	11
LRRC34	Esophagus_Mucosa	0.16	5.97	2.30E-09	0.02	0.04	2.73E-03	1.49E-03	rs9878797	-10.62	26	27	27
LRRC34	Heart_Atrial_Appendage	0.11	9.67	4.16E-22	0.10	0.06	1.42E-03	1.89E-03	rs1997392	-14.17	48	49	49
LRRC34	Heart_Left_Ventricle	0.24	12.46	1.24E-35	0.04	0.11	1.92E-06	4.41E-06	rs1997392	-14.17	13	14	14
LRRC34	Muscle_Skeletal	0.17	8.09	5.99E-16	0.03	0.06	1.07E-06	1.21E-06	rs6793160	-11.54	16	18	18
LRRC34	Nerve_Tibial	0.25	14.31	1.92E-46	0.04	0.10	2.21E-07	2.06E-07	rs3821383	-14.15	11	12	12
LRRC34	Thyroid	0.44	11.91	1.06E-32	0.01	0.04	1.40E-03	8.07E-04	rs6793160	-11.54	7	9	9
LRRIQ4	Cells_Transformed_fibroblasts	-0.41	-10.95	6.89E-28	0.01	0.01	1.32E-01	4.22E-02	rs6793160	-11.54	13	13	13
LRRIQ4	Testis	-0.21	-12.45	1.45E-35	0.05	0.09	1.74E-04	1.83E-04	rs1997392	-14.17	22	23	23
DCLK2	Spleen												
NAF1	Muscle_Skeletal	-0.20	-7.75	9.15E-15	0.02	0.06	5.47E-06	5.53E-06	rs1055263	7.79	3	3	3
NAF1	Skin_Sun_Exposed_Lower_leg	-0.22	-8.87	7.50E-19	0.02	0.02	3.01E-02	1.24E-02	rs1351222	9.44	28	28	28
NAF1	Thyroid	-0.11	-6.37	1.87E-10	0.04	0.03	3.80E-03	2.00E-03	rs1351222	9.44	24	28	28
NAF1	Cells_Transformed_fibroblasts												
RP11-563E2.2	Adipose_Subcutaneous												
RP11-563E2.2	Nerve_Tibial												
RP11-563E2.2	Skin_Not_Sun_Exposed_Suprapubic												
RP11-563E2.2	Skin_Sun_Exposed_Lower_leg												
PAPD4	Skin_Not_Sun_Exposed_Suprapubic												
PAPD4	Breast_Mammary_Tissue												
PAPD4	Colon_Transverse												
PAPD4	Esophagus_Muscularis												
PAPD4	Lung												
PAPD4	Nerve_Tibial												
PAPD4	Skin_Sun_Exposed_Lower_leg												
PAPD4	Small_Intestine_Terminal_Ileum												
PAPD4	Thyroid												
PSORS1C1	Heart_Left_Ventricle	0.03	5.40	6.52E-08	0.41	0.48	8.24E-29	3.56E-27	rs3020644	5.56	35	35	35
CCHCR1	Colon_Sigmoid	-0.04	-5.66	1.56E-08	0.21	0.13	2.87E-05	9.86E-05	rs707939	6.22	34	34	34
CCHCR1	Nerve_Tibial	-0.04	-5.20	1.98E-07	0.30	0.39	4.49E-29	3.03E-28	rs3094005	-5.08	26	27	27
POU5F1	Adipose_Visceral_Omentum	-0.05	-5.27	1.36E-07	0.17	0.18	2.10E-09	9.67E-09	rs3130484	-4.71	42	42	42
HLA-C	Adrenal_Gland	-0.04	-5.48	4.37E-08	0.31	0.56	9.29E-24	9.31E-22	rs2736428	5.79	38	39	39
HLA-C	Artery_Aorta	-0.03	-5.32	1.06E-07	0.52	0.72	1.07E-55	1.44E-52	rs2075800	6.04	99	100	100
HLA-C	Artery_Coronary	-0.04	-5.49	3.98E-08	0.31	0.49	7.31E-19	5.47E-17	rs3094005	-5.08	60	61	61
HLA-C	Artery_Tibial	-0.03	-5.45	5.06E-08	0.36	0.71	1.79E-78	6.72E-76	rs2844458	5.36	33	34	34
HLA-C	Breast_Mammary_Tissue	-0.03	-5.24	1.64E-07	0.37	0.61	6.04E-39	1.42E-36	rs3020644	5.56	33	34	34
HLA-C	Colon_Transverse	-0.03	-5.49	4.03E-08	0.46	0.63	2.57E-38	6.71E-36	rs3094005	-5.08	64	65	65
HLA-C	Heart_Atrial_Appendage	-0.03	-5.32	1.06E-07	0.37	0.52	3.78E-27	2.93E-25	rs3094005	-5.08	45	47	47
HLA-C	Heart_Left_Ventricle	-0.04	-5.24	1.60E-07	0.30	0.58	3.52E-37	4.97E-35	rs3094005	-5.08	36	37	37

Sentinel SNP and tissue-specific gene expression							Co-localisation				
SNP	CHR	Gene_start	Gene_end	Gene name	Gene nsnps	Tissue	H0_abf	H1_abf	H2_abf	H3_abf	H4_abf
rs2736176	6	31236526	31239882	HLA-C	17059	Pancreas	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17047	Spleen	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31236526	31239882	HLA-C	17058	Stomach	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31321649	31324219	HLA-B	17420	Liver	0.00	0.07	0.00	0.92	0.01
rs2736176	6	31588497	31605548	PRRC2A	18856	Whole_Blood	0.00	0.00	0.00	0.55	0.45
rs2736176	6	31606805	31620482	BAG6	19732	Artery_Tibial	0.00	0.00	0.00	1.00	0.00
rs2736176	6	31606805	31620482	BAG6	19698	Brain_Anterior_cingulate_cortex_BA24	0.00	0.00	0.00	0.02	0.98
rs2736176	6	31606805	31620482	BAG6	19577	Brain_Frontal_Cortex_BA9	0.00	0.00	0.00	0.01	0.99
rs2736176	6	31606805	31620482	BAG6	19701	Brain_Hypothalamus	0.00	0.00	0.00	0.01	0.99
rs2736176	6	31606805	31620482	BAG6	19732	Breast_Mammary_Tissue	0.00	0.30	0.00	0.45	0.24
rs2736176	6	31606805	31620482	BAG6	19732	Heart_Left_Ventricle	0.00	0.04	0.00	0.30	0.67
rs2736176	6	31606805	31620482	BAG6	19694	Ovary	0.00	0.03	0.00	0.09	0.88
rs2736176	6	31606805	31620482	BAG6	19730	Stomach	0.00	0.00	0.00	0.10	0.90
rs2736176	6	31629006	31634060	GPANK1	20119	Adipose_Subcutaneous	0.00	0.09	0.00	0.87	0.04
rs2736176	6	31633168	31637847	CSNK2B	20121	Adipose_Visceral_Omentum	0.00	0.04	0.00	0.11	0.86
rs2736176	6	31633168	31637847	CSNK2B	20121	Colon_Sigmoid	0.00	0.00	0.00	0.02	0.98
rs2736176	6	31633168	31637847	CSNK2B	20121	Esophagus_Muscularis	0.00	0.00	0.00	0.10	0.90
rs2736176	6	31633168	31637847	CSNK2B	20121	Heart_Atrial_Appendage	0.00	0.00	0.00	0.01	0.99
rs2736176	6	31633168	31637847	CSNK2B	20121	Heart_Left_Ventricle	0.00	0.00	0.00	0.05	0.95
rs2736176	6	31633168	31637847	CSNK2B	20121	Lung	0.00	0.00	0.00	0.03	0.97
rs2736176	6	31633168	31637847	CSNK2B	20121	Muscle_Skeletal	0.00	0.00	0.00	0.06	0.94
rs2736176	6	31847536	31865461	EHMT2	22103	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.00	0.30	0.70
rs2736176	6	31847536	31865461	EHMT2	22103	Testis	0.00	0.08	0.00	0.50	0.42
rs2736176	6	31865562	31913426	C2	22102	Whole_Blood	0.00	0.00	0.00	0.98	0.01
rs2736176	6	31865562	31913426	C2	22102	Testis	0.00	0.00	0.00	0.16	0.84
rs2736176	6	32083112	32096030	ATF6B	22268	Small_Intestine_Terminal_Ileum	0.00	0.03	0.00	0.12	0.85
rs2736176	6	31588497	31605548	PRRC2A	18856	Adipose_Subcutaneous	0.00	0.00	0.00	0.04	0.96
rs2736176	6	31588497	31605548	PRRC2A	18856	Muscle_Skeletal	0.00	0.00	0.00	0.15	0.85
rs2736176	6	31588497	31605548	PRRC2A	18856	Thyroid	0.00	0.00	0.00	0.10	0.90
rs2736176	6	31830969	31846823	SLC44A4	22041	Thyroid	0.00	0.00	0.00	0.11	0.89
rs59294613	7	124386051	124405681	GPR37	6504	Artery_Coronary	0.00	0.25	0.00	0.46	0.30
rs59294613	7	124570038	124819369	RP11-3B12.1	6950	Thyroid	0.00	0.00	0.00	0.05	0.95
rs59294613	7	124570038	124819369	RP11-3B12.1	6950	Pancreas	0.00	0.01	0.00	0.11	0.88
rs9419958	10	105642300	105677963	OBFC1	4964	Cells_Transformed_fibroblasts	0.00	0.00	0.00	0.07	0.93
rs9419958	10	105642300	105677963	OBFC1	4964	Colon_Transverse	0.00	0.00	0.00	0.03	0.97
rs9419958	10	105642300	105677963	OBFC1	4964	Esophagus_Gastroesophageal_Junction	0.00	0.11	0.00	0.08	0.82
rs9419958	10	105642300	105677963	OBFC1	4964	Esophagus_Mucosa	0.00	0.00	0.00	0.01	0.99
rs9419958	10	105642300	105677963	OBFC1	4964	Esophagus_Muscularis	0.00	0.01	0.00	0.02	0.97
rs9419958	10	105642300	105677963	OBFC1	4964	Heart_Atrial_Appendage	0.00	0.00	0.00	0.03	0.97
rs9419958	10	105642300	105677963	OBFC1	4964	Lung	0.00	0.03	0.00	0.04	0.94
rs9419958	10	105642300	105677963	OBFC1	4964	Skin_Sun_Exposed_Lower_leg	0.00	0.01	0.00	0.12	0.88
rs9419958	10	105642300	105677963	OBFC1	4964	Thyroid	0.00	0.11	0.00	0.06	0.83
rs9419958	10	105637132	105639519	RP11-541N10.3	4957	Adipose_Subcutaneous	0.00	0.03	0.00	0.02	0.95
rs9419958	10	105637132	105639519	RP11-541N10.3	4957	Thyroid	0.00	0.05	0.00	0.04	0.91
rs9419958	10	105726959	105788991	SLK	5037	Artery_Tibial	0.00	0.01	0.00	0.08	0.91
rs228595	11	108093211	108239829	ATM	6082	Cells_Transformed_fibroblasts	0.00	0.00	0.00	0.36	0.64
rs228595	11	107992478	108018503	ACAT1	6123	Artery_Aorta	0.00	0.00	0.00	0.15	0.84
rs2302588	14	73436159	73493920	ZFYVE1	6890	Colon_Sigmoid	0.00	0.05	0.00	0.09	0.86
rs3213718	14	90862846	90874605	CALM1	6373	Testis	0.00	0.00	0.04	0.02	0.93
rs12909131	15	50150435	50411654	ATP8B4	6190	Brain_Cortex	0.00	0.06	0.00	0.10	0.83
rs12909131	15	50171801	50175722	CTD-2647E9.3	6219	Lung	0.00	0.10	0.00	0.07	0.83
rs55710439	15	65204101	65251042	ANKDD1A	4785	Thyroid	0.00	0.00	0.01	0.02	0.97
rs11640926	16	1408901	1411406	LA16c-316G12.2	8048	Artery_Coronary	0.02	0.02	0.07	0.04	0.85
rs11640926	16	1256560	1257124	RP11-616M22.3	8297	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.00	0.09	0.06	0.85
rs3785074	16	69354043	69373332	COG8	4584	Brain_Cerebellum	0.00	0.63	0.00	0.30	0.08
rs3785074	16	69354043	69373332	COG8	4577	Brain_Cortex	0.00	0.00	0.00	0.66	0.34
rs3785074	16	69354043	69373332	COG8	4533	Brain_Frontal_Cortex_BA9	0.00	0.02	0.00	0.16	0.82
rs3785074	16	69354043	69373332	COG8	4586	Esophagus_Muscularis	0.00	0.02	0.00	0.08	0.90
rs3785074	16	69363900	69364498	PDF	4599	Adipose_Subcutaneous	0.00	0.00	0.00	0.04	0.96
rs3785074	16	69363900	69364498	PDF	4599	Adipose_Visceral_Omentum	0.00	0.00	0.00	0.89	0.11
rs3785074	16	69363900	69364498	PDF	4599	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.00	0.00	0.84	0.16
rs3785074	16	69363900	69364498	PDF	4599	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.00	0.73	0.27
rs3785074	16	69363900	69364498	PDF	4599	Artery_Tibial	0.00	0.00	0.00	0.18	0.82
rs3785074	16	69363900	69364498	PDF	4599	Breast_Mammary_Tissue	0.00	0.03	0.00	0.06	0.91
rs3785074	16	69363900	69364498	PDF	4599	Colon_Transverse	0.00	0.00	0.00	0.03	0.97
rs3785074	16	69363900	69364498	PDF	4599	Esophagus_Gastroesophageal_Junction	0.00	0.03	0.00	0.13	0.84
rs3785074	16	69363900	69364498	PDF	4599	Lung	0.00	0.00	0.00	0.15	0.85
rs3785074	16	69363900	69364498	PDF	4599	Nerve_Tibial	0.00	0.00	0.00	0.02	0.98
rs3785074	16	69363900	69364498	PDF	4599	Thyroid	0.00	0.00	0.00	0.16	0.84
rs3785074	16	69377151	69385712	TMED6	4549	Cells_EBV-transformed_lymphocytes	0.00	0.47	0.00	0.29	0.24

S-PrediXcan (if overlapped with colocalisation)													
Gene name	Tissue	effect size	zscore	pvalue	var_g	pred_perf			BEST_GWAS		n_snps		
						r^2	pval	qval	ID	Z	used,cov,mode		
HLA-C	Pancreas	-0.04	-5.37	7.87E-08	0.23	0.50	9.57E-24	4.39E-22	rs2075800	6.04	21	22	22
HLA-C	Spleen	-0.04	-5.78	7.27E-09	0.36	0.52	1.37E-15	8.47E-14	rs707939	6.22	49	50	50
HLA-C	Stomach	-0.04	-5.16	2.46E-07	0.23	0.59	4.26E-34	9.26E-32	rs2075800	6.04	20	21	21
HLA-B	Liver	-0.06	-5.77	8.00E-09	0.12	0.08	5.82E-03	1.24E-02	rs2736428	5.79	22	22	22
PRRC2A	Whole_Blood	0.13	5.31	1.10E-07	0.02	0.03	7.70E-04	5.10E-04	rs707939	6.22	13	13	13
BAG6	Artery_Tibial	0.05	5.77	8.15E-09	0.21	0.30	2.47E-23	1.19E-22	rs707939	6.22	31	31	31
BAG6	Brain_Anterior_cingulate_cortex_BA24	0.10	6.49	8.73E-11	0.05	0.09	9.04E-03	1.93E-02	rs707939	6.22	11	11	11
BAG6	Brain_Frontal_Cortex_BA9	0.06	7.01	2.42E-12	0.15	0.18	2.97E-05	1.42E-04	rs707939	6.22	21	21	21
BAG6	Brain_Hypothalamus	0.06	5.49	4.12E-08	0.12	0.10	3.55E-03	1.08E-02	rs707939	6.22	17	17	17
BAG6	Breast_Mammary_Tissue	0.20	5.89	3.95E-09	0.01	0.02	4.89E-02	4.06E-02	rs707939	6.22	2	2	2
BAG6	Heart_Left_Ventricle	0.11	5.43	5.61E-08	0.03	0.03	1.74E-02	1.49E-02	rs2736428	5.79	11	11	11
BAG6	Ovary	0.07	6.49	8.35E-11	0.11	0.10	3.77E-03	8.70E-03	rs2734325	5.37	19	19	19
BAG6	Stomach												
GPANK1	Adipose_Subcutaneous	0.17	5.19	2.13E-07	0.01	0.01	7.81E-02	3.03E-02	rs3094005	-5.08	10	10	10
CSNK2B	Adipose_Visceral_Omentum												
CSNK2B	Colon_Sigmoid												
CSNK2B	Esophagus_Muscularis												
CSNK2B	Heart_Atrial_Appendage												
CSNK2B	Heart_Left_Ventricle												
CSNK2B	Lung												
CSNK2B	Muscle_Skeletal												
EHMT2	Skin_Sun_Exposed_Lower_leg	-0.14	-5.57	2.59E-08	0.02	0.03	1.58E-03	8.82E-04	rs2075800	6.04	9	9	9
EHMT2	Testis	-0.15	-5.63	1.78E-08	0.02	0.04	1.39E-02	8.82E-03	rs2736428	5.79	6	6	6
C2	Whole_Blood	-0.20	-5.29	1.21E-07	0.01	0.01	8.76E-02	3.55E-02	rs497309	-4.74	6	6	6
C2	Testis												
ATF6B	Small_Intestine_Terminal_Ileum												
PRRC2A	Adipose_Subcutaneous												
PRRC2A	Muscle_Skeletal												
PRRC2A	Thyroid												
SLC44A4	Thyroid												
GPR37	Artery_Coronary	-0.14	-5.79	6.90E-09	0.02	0.04	2.93E-02	3.47E-02	rs2170352	-6.96	4	4	4
RP11-3B12.1	Thyroid												
RP11-3B12.1	Pancreas												
OBFC1	Cells_Transformed_fibroblasts												
OBFC1	Colon_Transverse												
OBFC1	Esophagus_Gastroesophageal_Junction												
OBFC1	Esophagus_Mucosa	-0.10	-6.34	2.34E-10	0.06	0.03	7.73E-03	3.79E-03	rs9419958	-8.92	36	38	38
OBFC1	Esophagus_Muscularis												
OBFC1	Heart_Atrial_Appendage												
OBFC1	Lung												
OBFC1	Skin_Sun_Exposed_Lower_leg	-0.11	-5.87	4.30E-09	0.04	0.04	2.67E-04	1.73E-04	rs9419958	-8.92	22	25	25
OBFC1	Thyroid												
RP11-541N10.3	Adipose_Subcutaneous												
RP11-541N10.3	Thyroid												
SLK	Artery_Tibial												
ATM	Cells_Transformed_fibroblasts	0.08	5.19	2.10E-07	0.05	0.11	3.40E-08	3.25E-08	rs7931930	-5.14	11	11	11
ACAT1	Artery_Aorta												
ZFYVE1	Colon_Sigmoid												
CALM1	Testis												
ATP8B4	Brain_Cortex												
CTD-2647E9.3	Lung												
ANKDD1A	Thyroid												
LA16c-316G12.2	Artery_Coronary												
RP11-616M22.3	Skin_Not_Sun_Exposed_Suprapubic												
COG8	Brain_Cerebellum	0.12	5.69	1.28E-08	0.03	0.04	3.90E-02	2.61E-02	rs3785073	6.19	7	7	7
COG8	Brain_Cortex	0.05	5.24	1.60E-07	0.11	0.10	1.68E-03	3.62E-03	rs877534	5.03	16	16	16
COG8	Brain_Frontal_Cortex_BA9	0.05	5.48	4.33E-08	0.15	0.12	5.64E-04	1.62E-03	rs877534	5.03	25	25	25
COG8	Esophagus_Muscularis												
PDF	Adipose_Subcutaneous	-0.18	-5.76	8.34E-09	0.01	0.04	1.06E-03	6.64E-04	rs12922774	5.91	3	3	3
PDF	Adipose_Visceral_Omentum	-0.11	-5.40	6.51E-08	0.03	0.05	1.63E-03	2.18E-03	rs7191614	6.15	9	9	9
PDF	Skin_Not_Sun_Exposed_Suprapubic	-0.09	-5.19	2.06E-07	0.04	0.05	1.39E-03	1.54E-03	rs12922774	5.91	16	16	16
PDF	Skin_Sun_Exposed_Lower_leg	-0.24	-5.90	3.56E-09	0.01	0.01	5.21E-02	2.01E-02	rs3785074	6.24	6	6	6
PDF	Artery_Tibial												
PDF	Breast_Mammary_Tissue												
PDF	Colon_Transverse												
PDF	Esophagus_Gastroesophageal_Junction												
PDF	Lung												
PDF	Nerve_Tibial												
PDF	Thyroid												
TMED6	Cells_EBV-transformed_lymphocytes	0.07	5.45	5.09E-08	0.08	0.11	2.68E-04	6.17E-04	rs3743669	6.08	14	14	14

Sentinel SNP and tissue-specific gene expression							Co-localisation				
SNP	CHR	Gene_start	Gene_end	Gene name	Gene nsnps	Tissue	H0_abf	H1_abf	H2_abf	H3_abf	H4_abf
rs3785074	16	69377151	69385712	TMED6	4560	Adipose_Subcutaneous	0.00	0.00	0.00	0.10	0.90
rs3785074	16	69389464	69442474	TERF2	4423	Brain_Cerebellar_Hemisphere	0.00	0.00	0.00	0.04	0.96
rs3785074	16	69389464	69442474	TERF2	4479	Brain_Cerebellum	0.00	0.01	0.00	0.02	0.97
rs3785074	16	69389464	69442474	TERF2	4482	Colon_Sigmoid	0.00	0.00	0.00	0.03	0.96
rs3785074	16	69389464	69442474	TERF2	4453	Ovary	0.00	0.01	0.00	0.02	0.98
rs3785074	16	69389464	69442474	TERF2	4482	Thyroid	0.00	0.00	0.00	0.01	0.99
rs3785074	16	69373571	69377014	NIP7	4585	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.07	0.00	0.12	0.81
rs3785074	16	69345259	69358945	VPS4A	4635	Muscle_Skeletal	0.00	0.07	0.00	0.11	0.82
rs62053580	16	74655292	74700779	RFWD3	6797	Adrenal_Gland	0.00	0.00	0.00	0.05	0.95
rs62053580	16	74655292	74700779	RFWD3	6771	Cells_EBV-transformed_lymphocytes	0.00	0.00	0.00	0.09	0.91
rs62053580	16	74655292	74700779	RFWD3	6801	Esophagus_Mucosa	0.00	0.00	0.00	0.07	0.93
rs62053580	16	74655292	74700779	RFWD3	6801	Muscle_Skeletal	0.00	0.00	0.00	0.19	0.81
rs62053580	16	74655292	74700779	RFWD3	6801	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.00	0.00	0.05	0.95
rs62053580	16	74655292	74700779	RFWD3	6801	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.00	0.07	0.93
rs62053580	16	74456018	74469152	RP11-252A24.5	6741	Brain_Cerebellum	0.00	0.02	0.00	0.06	0.92
rs62053580	16	74481325	74483790	RP11-252A24.7	6759	Artery_Tibial	0.00	0.03	0.00	0.13	0.83
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Adipose_Subcutaneous	0.00	0.00	0.00	0.04	0.96
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Adipose_Visceral_Omentum	0.00	0.00	0.00	0.60	0.40
rs7194734	16	82181403	82203831	MPHOSPH6	10987	Adrenal_Gland	0.00	0.00	0.00	0.19	0.81
rs7194734	16	82181403	82203831	MPHOSPH6	10986	Artery_Aorta	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10936	Artery_Coronary	0.00	0.08	0.00	0.14	0.78
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Artery_Tibial	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10949	Brain_Cerebellar_Hemisphere	0.00	0.00	0.00	0.09	0.91
rs7194734	16	82181403	82203831	MPHOSPH6	10976	Brain_Cerebellum	0.00	0.00	0.00	0.60	0.40
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Breast_Mammary_Tissue	0.00	0.00	0.00	0.06	0.94
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Colon_Transverse	0.00	0.00	0.00	0.13	0.87
rs7194734	16	82181403	82203831	MPHOSPH6	10986	Esophagus_Gastroesophageal_Junction	0.00	0.00	0.00	0.16	0.84
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Esophagus_Mucosa	0.00	0.01	0.00	0.99	0.00
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Esophagus_Muscularis	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Heart_Atrial_Appendage	0.00	0.00	0.00	0.05	0.95
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Heart_Left_Ventricle	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10976	Liver	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Lung	0.00	0.00	0.00	0.11	0.89
rs7194734	16	82181403	82203831	MPHOSPH6	10870	Minor_Salivary_Gland	0.00	0.01	0.00	0.04	0.94
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Muscle_Skeletal	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Nerve_Tibial	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Pancreas	0.00	0.00	0.00	0.04	0.96
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Skin_Not_Sun_Exposed_Suprapubic	0.00	0.00	0.00	0.04	0.96
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.00	0.04	0.96
rs7194734	16	82181403	82203831	MPHOSPH6	10948	Small_Intestine_Terminal_Ileum	0.00	0.00	0.00	0.05	0.95
rs7194734	16	82181403	82203831	MPHOSPH6	10982	Spleen	0.00	0.00	0.00	0.60	0.40
rs7194734	16	82181403	82203831	MPHOSPH6	10991	Stomach	0.00	0.00	0.00	0.03	0.97
rs7194734	16	82181403	82203831	MPHOSPH6	10991	Testis	0.00	0.00	0.00	0.02	0.98
rs7194734	16	82181403	82203831	MPHOSPH6	10992	Thyroid	0.00	0.00	0.00	0.03	0.97
rs144204502	17	76170160	76183314	TK1	8131	Brain_Nucleus_accumbens_basal_ganglia	0.00	0.00	0.01	0.01	0.99
rs144204502	17	76170160	76183314	TK1	8142	Esophagus_Mucosa	0.02	0.01	0.01	0.01	0.95
rs144204502	17	76170160	76183314	TK1	8126	Brain_Putamen_basal_ganglia	0.01	0.00	0.03	0.02	0.94
rs2124616	18	657604	673578	TYMS	6126	Cells_EBV-transformed_lymphocytes	0.00	0.00	0.00	0.03	0.96
rs8105767	19	22235254	22274282	ZNF257	8158	Artery_Coronary	0.00	0.00	0.00	0.08	0.92
rs8105767	19	22235254	22274282	ZNF257	8160	Brain_Cerebellum	0.00	0.00	0.00	0.03	0.97
rs8105767	19	22235254	22274282	ZNF257	8110	Brain_Hypothalamus	0.00	0.00	0.00	0.06	0.94
rs8105767	19	22235254	22274282	ZNF257	8105	Brain_Spinal_cord_cervical_c-1	0.00	0.00	0.00	0.15	0.84
rs8105767	19	22235254	22274282	ZNF257	8163	Colon_Transverse	0.00	0.00	0.00	0.00	1.00
rs8105767	19	22235254	22274282	ZNF257	8164	Esophagus_Mucosa	0.00	0.00	0.00	0.00	1.00
rs8105767	19	22235254	22274282	ZNF257	8164	Nerve_Tibial	0.00	0.00	0.00	0.55	0.45
rs8105767	19	22235254	22274282	ZNF257	8157	Prostate	0.00	0.00	0.00	0.02	0.98
rs8105767	19	22235254	22274282	ZNF257	8164	Skin_Sun_Exposed_Lower_leg	0.00	0.00	0.00	0.19	0.81
rs8105767	19	22235254	22274282	ZNF257	8158	Small_Intestine_Terminal_Ileum	0.00	0.00	0.00	0.16	0.84
rs8105767	19	22235254	22274282	ZNF257	8164	Stomach	0.00	0.00	0.00	0.05	0.95
rs8105767	19	22235254	22274282	ZNF257	8164	Whole_Blood	0.00	0.00	0.00	0.04	0.96
rs8105767	19	22361893	22379753	ZNF676	8116	Lung	0.00	0.01	0.00	0.04	0.95
rs1744757	20	35624752	35724187	RBL1	3897	Muscle_Skeletal	0.00	0.00	0.00	0.09	0.91
rs1744757	20	35624752	35724187	RBL1	3897	Cells_Transformed_fibroblasts	0.00	0.03	0.00	0.12	0.85
rs1744757	20	35518632	35580246	SAMHD1	3834	Whole_Blood	0.00	0.00	0.00	0.05	0.95
rs1744757	20	35518632	35580246	SAMHD1	3819	Ovary	0.00	0.07	0.00	0.09	0.84
rs75691080	20	62585007	62585495	AL118506.1	5283	Artery_Tibial	0.00	0.32	0.00	0.64	0.04
rs75691080	20	62704534	62711323	RGS19	4672	Brain_Hypothalamus	0.00	0.69	0.00	0.27	0.04

S-PrediXcan (if overlapped with colocalisation)													
Gene name	Tissue	effect size	zscore	pvalue	var_g	pred_perf			BEST_GWAS		n_snps		
						r^2	pval	qval	ID	Z	used	cov	mode
TMED6	Adipose_Subcutaneous												
TERF2	Brain_Cerebellar_Hemisphere												
TERF2	Brain_Cerebellum												
TERF2	Colon_Sigmoid												
TERF2	Ovary	-0.08	-5.25	1.50E-07	0.06	0.09	4.66E-03	1.03E-02	rs3785074	6.24	9	9	9
TERF2	Thyroid	-0.08	-5.76	8.48E-09	0.08	0.15	1.92E-11	2.93E-11	rs3785074	6.24	7	8	8
NIP7	Skin_Not_Sun_Exposed_Suprapubic												
VPS4A	Muscle_Skeletal												
RFWD3	Adrenal_Gland												
RFWD3	Cells_EBV-transformed_lymphocytes												
RFWD3	Esophagus_Mucosa												
RFWD3	Muscle_Skeletal												
RFWD3	Skin_Not_Sun_Exposed_Suprapubic												
RFWD3	Skin_Sun_Exposed_Lower_leg												
RP11-252A24.5	Brain_Cerebellum												
RP11-252A24.7	Artery_Tibial												
MPHOSPH6	Adipose_Subcutaneous												
MPHOSPH6	Adipose_Visceral_Omentum	0.06	5.43	5.61E-08	0.12	0.18	1.79E-09	8.27E-09	rs2967355	-6.11	33	33	33
MPHOSPH6	Adrenal_Gland												
MPHOSPH6	Artery_Aorta												
MPHOSPH6	Artery_Coronary	0.09	5.39	7.19E-08	0.06	0.10	6.36E-04	1.67E-03	rs2967355	-6.11	22	22	22
MPHOSPH6	Artery_Tibial	0.05	5.78	7.62E-09	0.18	0.30	2.68E-23	1.28E-22	rs2967374	-6.13	21	21	21
MPHOSPH6	Brain_Cerebellar_Hemisphere												
MPHOSPH6	Brain_Cerebellum	0.04	5.36	8.20E-08	0.21	0.35	6.72E-11	4.69E-10	rs2967355	-6.11	17	17	17
MPHOSPH6	Breast_Mammary_Tissue	0.20	5.95	2.64E-09	0.01	0.05	3.74E-03	5.08E-03	rs2967374	-6.13	10	11	11
MPHOSPH6	Colon_Transverse												
MPHOSPH6	Esophagus_Gastroesophageal_Junction												
MPHOSPH6	Esophagus_Mucosa	0.08	5.34	9.49E-08	0.06	0.11	8.51E-08	1.01E-07	rs12102917	-5.16	16	16	16
MPHOSPH6	Esophagus_Muscularis	0.05	5.32	1.02E-07	0.17	0.27	2.39E-16	1.02E-15	rs2967374	-6.13	34	34	34
MPHOSPH6	Heart_Atrial_Appendage												
MPHOSPH6	Heart_Left_Ventricle	0.05	5.98	2.28E-09	0.19	0.34	1.33E-18	1.75E-17	rs2967374	-6.13	11	11	11
MPHOSPH6	Liver	0.06	5.89	3.79E-09	0.13	0.31	4.14E-09	6.79E-08	rs7202258	-6.09	13	13	13
MPHOSPH6	Lung	0.09	5.76	8.23E-09	0.06	0.14	1.11E-10	2.76E-10	rs2911423	-5.43	21	21	21
MPHOSPH6	Minor_Salivary_Gland												
MPHOSPH6	Muscle_Skeletal	0.04	5.51	3.64E-08	0.21	0.32	3.48E-32	3.13E-31	rs2967355	-6.11	22	22	22
MPHOSPH6	Nerve_Tibial	0.04	5.16	2.46E-07	0.22	0.28	5.96E-20	1.94E-19	rs2967355	-6.11	51	51	51
MPHOSPH6	Pancreas	0.08	5.83	5.51E-09	0.09	0.15	1.41E-06	3.68E-06	rs2967374	-6.13	21	21	21
MPHOSPH6	Skin_Not_Sun_Exposed_Suprapubic												
MPHOSPH6	Skin_Sun_Exposed_Lower_leg												
MPHOSPH6	Small_Intestine_Terminal_Ileum												
MPHOSPH6	Spleen												
MPHOSPH6	Stomach	0.04	5.84	5.36E-09	0.28	0.35	1.17E-17	2.62E-16	rs2967374	-6.13	35	35	35
MPHOSPH6	Testis	0.04	5.28	1.29E-07	0.23	0.34	1.44E-15	8.48E-15	rs2967374	-6.13	39	39	39
MPHOSPH6	Thyroid	0.04	5.90	3.57E-09	0.23	0.35	1.00E-27	5.68E-27	rs2967374	-6.13	22	22	22
TK1	Brain_Nucleus_accumbens_basal_ganglia												
TK1	Esophagus_Mucosa												
TK1	Brain_Putamen_basal_ganglia												
TYMS	Cells_EBV-transformed_lymphocytes												
ZNF257	Artery_Coronary												
ZNF257	Brain_Cerebellum												
ZNF257	Brain_Hypothalamus												
ZNF257	Brain_Spinal_cord_cervical_c-1												
ZNF257	Colon_Transverse	0.06	6.84	7.73E-12	0.14	0.22	1.44E-10	7.83E-10	rs8105767	7.22	19	19	19
ZNF257	Esophagus_Mucosa												
ZNF257	Nerve_Tibial	0.04	5.39	6.93E-08	0.23	0.30	1.77E-21	6.58E-21	rs8105767	7.22	33	33	33
ZNF257	Prostate												
ZNF257	Skin_Sun_Exposed_Lower_leg	0.04	5.47	4.63E-08	0.21	0.25	2.50E-20	9.39E-20	rs8105767	7.22	40	41	41
ZNF257	Small_Intestine_Terminal_Ileum												
ZNF257	Stomach												
ZNF257	Whole_Blood	0.10	6.14	8.48E-10	0.05	0.07	1.17E-06	1.21E-06	rs1912576	6.63	13	15	15
ZNF676	Lung	0.21	7.02	2.20E-12	0.01	0.03	3.25E-03	2.49E-03	rs8105767	7.22	3	3	3
RBL1	Muscle_Skeletal												
RBL1	Cells_Transformed_fibroblasts												
SAMHD1	Whole_Blood												
SAMHD1	Ovary												
AL118506.1	Artery_Tibial	-0.45	-6.32	2.61E-10	0.00	0.01	8.30E-02	3.29E-02	rs2281929	6.32	2	2	2
RGS19	Brain_Hypothalamus	0.53	5.92	3.20E-09	0.00	0.12	1.47E-03	5.59E-03	rs6089956	5.58	3	3	3

Sentinel SNP and tissue-specific gene expression							Co-localisation				
SNP	CHR	Gene_start	Gene_end	Gene name	Gene nsnps	Tissue	H0_abf	H1_abf	H2_abf	H3_abf	H4_abf
rs75691080	20	62289163	62327606	RTEL1	6542	Breast_Mammary_Tissue	0.00	0.00	0.00	0.11	0.89
rs75691080	20	62289163	62327606	RTEL1	6542	Muscle_Skeletal	0.00	0.00	0.00	0.10	0.90
rs75691080	20	62289163	62327606	RTEL1	6542	Heart_Atrial_Appendage	0.00	0.06	0.00	0.05	0.89
rs75691080	20	62289163	62327606	RTEL1	6542	Adipose_Visceral_Omentum	0.00	0.00	0.00	0.04	0.96
rs75691080	20	62271061	62284780	STMN3	6556	Artery_Aorta	0.00	0.00	0.00	0.02	0.98
rs75691080	20	62271061	62284780	STMN3	6555	Colon_Sigmoid	0.00	0.07	0.00	0.05	0.88
rs75691080	20	62271061	62284780	STMN3	6556	Cells_Transformed_fibroblasts	0.00	0.00	0.00	0.01	0.99
rs75691080	20	62271061	62284780	STMN3	6556	Artery_Tibial	0.00	0.00	0.00	0.01	0.99
rs75691080	20	62328021	62329995	TNFRSF6B	6353	Adipose_Visceral_Omentum	0.00	0.03	0.00	0.05	0.92
rs75691080	20	62328021	62329995	TNFRSF6B	6353	Cells_Transformed_fibroblasts	0.00	0.03	0.00	0.04	0.93

S-PrediXcan (if overlapped with colocalisation)											
Gene name	Tissue	effect size	zscore	pvalue	var_g	pred_perf			BEST_GWAS		n_snps
						r^2	pval	qval	ID	Z	used,cov,mode
RTEL1	Breast_Mammary_Tissue										
RTEL1	Muscle_Skeletal										
RTEL1	Heart_Atrial_Appendage										
RTEL1	Adipose_Visceral_Omentum										
STMN3	Artery_Aorta										
STMN3	Colon_Sigmoid										
STMN3	Cells_Transformed_fibroblasts										
STMN3	Artery_Tibial										
TNFRSF6B	Adipose_Visceral_Omentum										
TNFRSF6B	Cells_Transformed_fibroblasts										

Supplementary Table 8: Integrated scoring of non-coding variants. Scoring was performed with SNP Nexus IW scoring tool.

Lead	SNP	CADD PHRED	deepseq sig_log^2	eigen score	eigen pc_score	fathmm nc_score	fitcons score	funseq score	gwava region	gwava tss	gwava unmatched	remm score	score integrated (11scores)	*p* -value	score integrated (10scores)	*p* -value
rs144204502	rs144204502	10.52	8.95	1.92	8.51	0.23	0.05	2.72	0.46	0.4	0.96	0.91	5.82	4.40E-03	6.10	2.68E-03
rs3213718	rs2300496	12.82	9.60	1.89	5.99	0.83	0.21	2.10	0.38	0.24	0.84	0.93	5.75	4.86E-03	5.82	3.98E-03
rs59192843	rs73301475	18.64	6.98	1.60	0.41	0.91	0.10	1.57	0.75	0.73	0.98	0.98	5.65	5.51E-03	5.81	4.04E-03
rs59294613	rs2239532	8.359	9.19	1.50	3.44	0.38	0.11	2.53	0.4	0.62	0.93	0.99	4.93	1.33E-02	5.10	1.00E-02
rs2736176	rs805299	8.595	6.72	1.68	1.43	0.93	0.14	2.28	0.57	0.42	0.87	0.96	4.77	1.60E-02	4.88	1.30E-02
rs3213718	rs12885713	9.24	6.15	0.63	2.77	0.99	0.03	1.48	0.77	0.79	0.97	0.81	4.51	2.13E-02	4.75	1.53E-02
rs3219104	rs907187	9.385	5.67	0.77	NA	0.67	0.09	3.20	0.63	0.46	0.99	0.88	4.62	1.88E-02	4.73	1.55E-02
rs10936600	rs10936599	11.71	9.08	1.79	0.76	0.95	0.66	NA	0.4	0.34	0.79	0.96	4.91	1.37E-02	4.60	1.81E-02
rs3785074	rs9939705	14.76	9.12	1.20	2.99	0.03	0.12	2.46	0.55	0.58	0.56	0.21	3.70	4.85E-02	3.85	3.99E-02
rs59192843	rs17094157	6.374	7.09	0.97	1.59	0.79	0.05	1.65	0.53	0.46	0.96	0.83	3.62	5.21E-02	3.80	4.18E-02
rs34991172	rs913455	18.28	8.30	1.26	0.05	0.85	0.71	NA	0.49	0.55	0.68	0.31	4.24	2.85E-02	3.68	4.70E-02
rs55710439	rs57438358	8.645	5.94	0.97	0.00	0.98	0.36	0.98	0.54	0.84	0.88	0.82	3.72	4.73E-02	3.63	4.97E-02

Supplementary Table 9: Identification of meQTLs. Independent SNPs associated with LTL at FDR<0.05 and their proxies (r^2<0.8) were searched in meQTL databases using PhenoScanner (section 2.2.6.3). Best proxy SNPs were those that exhibited the highest LD r^2 with locus sentinel SNPs; the corresponding lines indicate their associations with DNA methylation markers. Most significant meQTLs indicate SNPs that were most significantly associated with DNA methylation markers within each independent LTL signal, and their blocks show their associations with the DNA methylation markers and LD r^2 with the independent LTL signal SNPs.

Telomere GWAS Locus						Best Proxy SNP										
Chr	Region_Start	Region_End	lead_snp	nearest_gene	top_gene_candidate	rsID	hg19_coordinates	A1	A2	distance	r^2	correlated_alleles	beta	se	p	direction
1	113578755	114578755	rs12065882	MAGI3	AP4B1	rs12065882	chr1:114078755	A	G	0	1.00	A=A,G=G	NA	NA	3.92E-12	-
1	226062621	227062621	rs3219104	PARP1	PARP1	rs2377312	chr1:226561761	G	C	-860	1.00	A=G,C=C	NA	NA	2.57E-237	+
1	226062621	227062621	rs3219104	PARP1	PARP1	rs2377312	chr1:226561761	G	C	-860	1.00	A=G,C=C	NA	NA	7.85E-37	-
1	226062621	227062621	rs3219104	PARP1	PARP1	rs2377312	chr1:226561761	G	C	-860	1.00	A=G,C=C	NA	NA	2.57E-190	+
3	169014585	170014585	rs10936600	LRRC34	TERC,LRRC34	rs10936600	chr3:169514585	A	T	0	1.00	A=A,T=T	NA	NA	2.15E-88	-
4	78425743	79425743	rs62365174	PAPD4	PAPD4	rs62365174	chr5:78925743	A	G	0	1.00	A=A,G=G	NA	NA	5.64E-12	-
4	78425743	79425743	rs62365174	PAPD4	PAPD4	rs59421001	chr5:78925953	A	G	210	1.00	A=G,G=A	1.06	0.16	1.06E-10	+
11	107605593	108605593	rs228595	ATM	ATM	rs228595	chr11:108105593	A	G	0	1.00	G=G,A=A	-0.23	0.04	8.37E-10	-
12	13930807	14930807	rs112655343	ATF7IP	ATF7IP	rs112655343	chr12:14430807	C	T	0	1.00	C=C,T=T	NA	NA	5.71E-26	-
14	72904752	73904752	rs2302588	DCAF4	DCAF4	rs78044039	chr14:73454645	A	G	49893	0.90	G=G,C=A	NA	NA	1.45E-12	+
14	72904752	73904752	rs2302588	DCAF4	DCAF4	rs76891117	chr14:73399837	A	G	-4915	1.00	G=A,C=G	NA	NA	3.04E-124	-
14	74014120	75014120	rs59192843	BBOF1	ENTPD5	rs140682464	chr14:74520830	C	T	6710	0.82	T=C,G=T	NA	NA	5.35E-66	-
15	49887678	50887678	rs12909131	ATP8B4	ATP8B4	rs12909131	chr15:50387678	C	T	0	1.00	C=C,T=T	NA	NA	7.29E-254	-
15	49887678	50887678	rs12909131	ATP8B4	ATP8B4	rs12909131	chr15:50387678	C	T	0	1.00	C=C,T=T	NA	NA	1.76E-65	-
15	49887678	50887678	rs12909131	ATP8B4	ATP8B4	rs12909131	chr15:50387678	C	T	0	1.00	C=C,T=T	-0.73	0.04	4.48E-54	-
16	68906986	69906986	rs3785074	TERF2	TERF2	rs3785074	chr16:69406986	A	G	0	1.00	A=A,G=G	NA	NA	1.54E-267	-
16	81699980	82699980	rs7194734	MPHOSPH6	MPHOSPH6	rs7194734	chr16:82199980	C	T	0	1.00	C=C,T=T	NA	NA	2.60E-40	+
16	81699980	82699980	rs7194734	MPHOSPH6	MPHOSPH6	rs7194734	chr16:82199980	C	T	0	1.00	C=C,T=T	NA	NA	4.83E-256	+
19	21715441	22715441	rs8105767	ZNF208	ZNF257	rs8105767	chr19:22215441	A	G	0	1.00	A=A,G=G	NA	NA	4.61E-22	-
20	61769750	62769750	rs73624724	ZBTB46	ZBTB46	rs73624724	chr20:62436398	C	T	0	1.00	T=T,C=C	NA	NA	6.71E-235	+

Extra columns are shown on the next page

Telomere GWAS Locus			Most significant methQTL for each region				DNA methylation marker and reference genes						
Chr	Region_Start	Region_End	rsID	hg19_coordinates	P_value	r^2	marker	RefGene_Name	RefGene_Group	marker_position	DMR	Enhancer	DHS
1	113578755	114578755	rs11588901	chr1:114091058	3.88E-13	0.86	cg16515600	RSBN1	Body	NA	NA	TRUE	NA
1	226062621	227062621	rs76887998	chr1:226539353	2.34E-251	0.92	cg04208928	LIN9	TSS1500	S_Shore	NA	NA	NA
1	226062621	227062621	rs4653729	chr1:226537535	6.80E-40	0.92	cg13952899	C1orf95	Body	S_Shore	RDMR	NA	NA
1	226062621	227062621	rs76887998	chr1:226539353	1.65E-193	0.92	cg23712594	PARP1	Body	N_Shelf	NA	TRUE	NA
3	169014585	170014585	rs9822885	chr3:169486144	6.49E-101	0.93	cg14222479	ARPM1	1stExon;5'UTR	S_Shore	NA	NA	TRUE
4	78425743	79425743	rs62365229	chr5:78958549	1.24E-17	0.94	cg02754494	HOMER1	TSS1500	Island	NA	NA	NA
4	78425743	79425743	rs62364124	chr5:78910132	2.75E-13	0.94	Percent-splice-in	PAPD4	protein_coding	NA	NA	NA	NA
11	107605593	108605593	rs11212620	chr11:108290959	2.36E-10	1.00	cg05081395	KDELC2	3'UTR	NA	NA	NA	NA
12	13930807	14930807	rs73056729	chr12:14432076	2.70E-26	0.95	cg19789919	ATF7IP	TSS200	Island	NA	NA	NA
14	72904752	73904752	rs362408	chr14:73698548	4.42E-16	0.90	cg19585100	PAPLN	TSS1500	N_Shore	RDMR	NA	NA
14	72904752	73904752	rs77694099	chr14:73398446	2.01E-124	1.00	cg23196123	DCAF4	TSS200	Island	NA	NA	NA
14	74014120	75014120	rs140682464	chr14:74520830	5.35E-66	0.82	cg18638434	C14orf45;ENTPD5	Body;TSS1500	S_Shore	NA	NA	NA
15	49887678	50887678	rs12903325	chr15:50353277	1.35E-263	0.94	cg00868652	ATP8B4	TSS200	NA	NA	TRUE	NA
15	49887678	50887678	rs41362650	chr15:50369375	4.69E-66	1.00	cg02726943	SLC27A2	1stExon	Island	NA	TRUE	TRUE
15	49887678	50887678	rs7172615	chr15:50357743	3.07E-57	0.97	cg23504246	C15orf33;FGF7	Body;TSS200	NA	NA	TRUE	NA
16	68906986	69906986	rs9939870	chr16:69396585	2.98E-268	1.00	cg02192472	PDF;COG8	TSS1500;Body	S_Shore	NA	NA	NA
16	81699980	82699980	rs2967352	chr16:82196676	8.91E-42	0.97	cg00540449	MPHOSPH6	TSS1500	S_Shore	NA	NA	NA
16	81699980	82699980	rs7203990	chr16:82185320	1.42E-259	0.97	cg19807685	HSD17B2	5'UTR;1stExon	NA	NA	NA	NA
19	21715441	22715441	rs7248898	chr19:22245354	3.58E-23	0.98	cg06852575	ZNF257	Body	NA	DMR	NA	TRUE
20	61769750	62769750	rs6011173	chr20:62460780	6.77E-237	1.00	cg01209296	ZBTB46	5'UTR	S_Shore	NA	NA	NA

Extra columns are shown on the next page

Telomere GWAS Locus			methQTL source and sample size					
Chr	Region_Start	Region_End	dataset	pmid	ancestry	year	tissue	n
1	113578755	114578755	BIOSQTL	27918535	European	2017	Whole blood	3841
1	226062621	227062621	BIOSQTL	27918535	European	2017	Whole blood	3841
1	226062621	227062621	BIOSQTL	27918535	European	2017	Whole blood	3841
1	226062621	227062621	BIOSQTL	27918535	European	2017	Whole blood	3841
3	169014585	170014585	BIOSQTL	27918535	European	2017	Whole blood	3841
4	78425743	79425743	BIOSQTL	27918535	European	2017	Whole blood	3841
4	78425743	79425743	BLUEPRINT	27863251	European	2016	Monocytes	194
11	107605593	108605593	Gaunt T	27036880	European	2016	Whole blood	837
12	13930807	14930807	BIOSQTL	27918535	European	2017	Whole blood	3841
14	72904752	73904752	BIOSQTL	27918535	European	2017	Whole blood	3841
14	72904752	73904752	BIOSQTL	27918535	European	2017	Whole blood	3841
14	74014120	75014120	BIOSQTL	27918535	European	2017	Whole blood	3841
15	49887678	50887678	BIOSQTL	27918535	European	2017	Whole blood	3841
15	49887678	50887678	BIOSQTL	27918535	European	2017	Whole blood	3841
15	49887678	50887678	Gaunt T	27036880	European	2016	Whole blood	837
16	68906986	69906986	BIOSQTL	27918535	European	2017	Whole blood	3841
16	81699980	82699980	BIOSQTL	27918535	European	2017	Whole blood	3841
16	81699980	82699980	BIOSQTL	27918535	European	2017	Whole blood	3841
19	21715441	22715441	BIOSQTL	27918535	European	2017	Whole blood	3841
20	61769750	62769750	BIOSQTL	27918535	European	2017	Whole blood	3841

Supplementary Table 10: Gene prioritisation. Evidence to support likely-causal genes, including nonsynonymous variants, eQTLs, known roles in telomere regulation and having other supportive information from literature. Genes were prioritised based on the most lines of evidence or on strength of evidence (deleteriously predicted mutations, known role in telomere biology and eQTLs in multiple tissues over single tissue).

Locus	Chr	bp	lead	Closest gene	# SNPs in LD	Nonsynonymous SNP	eQTL (S-PrediXcan and COLOC)	strong in COLOC only	Known biology	Other literature evidence	Prioritised gene(s)
Chr1p13.2	1	114078755	rs12065882	*MAGI3*	3		*AP4B1*	*PTPN22, AP4B1-AS1*			*AP4B1*
Chr1q24.2	1	167399643	rs35675808	*CD247*	0						
Chr1q42.12	1	226562621	rs3219104	*PARP1*	43	*PARP1**	*PARP1*	*C1orf95*	*PARP1*		*PARP1*
Chr2p16.2	2	54482703	rs754017156	*ACYP2*	0		*TSPYL6*				*TSPYL6*
Chr2q34	2	210663697	rs56810761	*UNC80*	0			*SNA1P1*			*SNA1P1*
Chr3q12.3	3	101232093	rs55749605	*SENP7*	76	*SENP7*				*SENP7*	*SENP7*
Chr3q13.2	3	112847045	rs2613954	*RP11-572M11.4*	21						
Chr3q26.2	3	169514585	rs10936600	*LRRC34*	47	*LRRC34*	*MYNN*		*TERC*		*TERC, LRRC34*
Chr4q13.3	4	71774347	rs13137667	*MOB1B*	49						
Chr4q31.23	4	151000830	rs60160057	*DCLK2*	64						
Chr4q32.2	4	164048199	rs4691895	*NAF1*	69	*NAF1*	*NAF1*		*NAF1*		*NAF1*
Chr5p15.33	5	1285974	rs7705526	*TERT*	0				*TERT*		*TERT*
Chr5p15.33	5	1287194	rs2853677	*TERT*	0				*TERT*		*TERT*
Chr5q14.1	5	78925743	rs62365174	*PAPD4*	137			*PAPD4*			*PAPD4*
Chr5q31.2	5	138964816	rs112347796	*UBE2D2*	0						
Chr6p22.2	6	25480328	rs34991172	*CARMIL1*	10						
Chr6p21.33	6	31587561	rs2736176	*PRRC2A*	11		*BAG6*	*CSNK2B, PRRC2A*		*CSNK2B, BAG6*	*CSNK2B, BAG6*
Chr7q31.33	7	124554267	rs59294613	*POT1*	118			*RP11-3B12.1*	*POT1*		*POT1*
Chr8p23.2	8	2882469	rs57415150	*CSMD1*	51						
Chr8q22.2	8	100917632	rs201375979	*COX6C*	3						
Chr10p15.1	10	5702259	rs2386642	*ASB13*	5						
Chr10q24.33	10	105675946	rs9419958	*STN1 (OBFC1)*	4		*STN1*	*RP11-541N10.3, SLK*	*STN1*		*STN1*

Extra rows are shown on the next page

Locus	Chr	bp	lead	Closest gene	# SNPs in LD	Nonsynonymous SNP	eQTL (S-PrediXcan and COLOC)	strong in COLOC only	Known biology	Other literature evidence	Prioritised gene(s)
Chr11q21	11	93404608	rs117037102	*CEP295*	10	*CEP295*					*CEP295*
Chr11q22.3	11	108105593	rs228595	*ATM*	57				*ATM*		*ATM*
Chr12p13.1	12	14430807	rs112655343	*ATF7IP*	2					*ATF7IP*	*ATF7IP*
Chr12q13.13	12	54592103	rs7311314	*SMUG1*	6					*SMUG1*	*SMUG1*
Chr14q24.2	14	73404752	rs2302588	*DCAF4*	71	*DCAF4, ZFYVE1*				*DCAF4*	*DCAF4*
Chr14q24.3	14	74514120	rs59192843	*CCDC176*	155						
Chr14q32.11	14	90869913	rs3213718	*CALM1*	7			*CALM1*			*CALM1*
Chr14q32.33	14	105494403	rs117536281	*CDCA4*	1						
Chr15q14	15	38930961	rs9972513	*RP11-275I4.2*	1						
Chr15q21.2	15	50387678	rs12909131	*ATP8B4*	16						
Chr15q21.3	15	55105443	rs117610974	*UNC13C*	0						
Chr15q22.31	15	65229816	rs55710439	*ANKDD1A*	47			*ANKDD1A*			*ANKDD1A*
Chr16p13.3	16	1249877	rs11640926	*CACNA1H*	0						
Chr16q22.1	16	69406986	rs3785074	*TERF2*	21		*TERF2, PDF, COG8*	*TMED6,*	*TERF2*		*TERF2*
Chr16q23.1	16	74680074	rs62053580	*RFWD3*	1			*RFWD3, RP11-252A24.5*		*RFWD3*	*RFWD3*
Chr16q23.3	16	82199980	rs7194734	*MPHOSPH6*	67	*MPHOSPH6*	*MPHOSPH6*				*MPHOSPH6*
Chr17q25.3	17	76183233	rs144204502	*TK1*	1			*TK1*			*TK1*
Chr18p11.32	18	661917	rs2124616	*TYMS*	8			*TYMS*			*TYMS*
Chr19p13.3	19	3939249	rs143276018	*NMRK2*	7						
Chr19p12	19	22215441	rs8105767	*ZNF257*	9		*ZNF257, ZNF676*				*ZNF257*
Chr19q13.2	19	39768216	rs11665818	*IFNL2*	0						
Chr20p12.3	20	5310273	rs6107615	*PROKR2*	1						
Chr20p12.3	20	7402809	rs6038821	*LINC01706*	19						
Chr20q11.23	20	35734863	rs1744757	*MROH8*	80			*SAMHD1, RBL1*			*SAMHD1, RBL1*
Chr20q13.33	20	62269750	rs75691080	*STMN3*	3			*RTEL1, STMN3, TNFRSF6B*	*RTEL1*		*RTEL1,STMN3*
Chr20q13.33	20	62291599	rs34978822	*RTEL1*	6	*RTEL1*			*RTEL1*		*RTEL1*
Chr20q13.33	20	62380527	rs932827	*ZBTB46*	1				*RTEL1*		*RTEL1*
Chr20q13.33	20	62436398	rs73624724	*ZBTB46*	112	*ZBTB46*			*RTEL1*		*ZBTB46*
Chr21q22.3	21	45994841	rs7276273	*KRTAP10-4*	1	*KRTAP10-4***					*KRTAP10-4*
Chr22q13.31	22	44698803	rs7510583	*KIAA1644*	0						

Supplementary Table 11: Pathway analysis. Prioritized genes or the closest genes to locus sentinel variants where no prioritization was possible were used as input to PANTHER (section 2.2.8.1). A statistical over-representation analysis was performed. Pathways over-represented at FDR<0.05 are shown.

GO biological process complete	Total gene counts	Observed count	expected %	fold Enrichment	*p*-value	FDR
regulation of telomeric loop disassembly (GO:1904533)	3	2	0.01	>100	5.29E-05	2.39E-02
regulation of single strand break repair (GO:1903516)	3	2	0.01	>100	5.29E-05	2.32E-02
negative regulation of t-circle formation (GO:1904430)	4	2	0.01	>100	7.92E-05	2.91E-02
establishment of protein localization to telomere (GO:0070200)	7	3	0.02	>100	1.41E-06	1.85E-03
telomeric loop disassembly (GO:0090657)	11	3	0.03	>100	4.24E-06	4.18E-03
protein localization to chromosome, telomeric region (GO:0070198)	15	4	0.04	>100	9.80E-08	2.21E-04
establishment of protein localization to chromosome (GO:0070199)	15	3	0.04	85.7	9.44E-06	7.10E-03
negative regulation of cellular senescence (GO:2000773)	19	3	0.04	67.66	1.77E-05	1.22E-02
negative regulation of telomere maintenance (GO:0032205)	38	6	0.09	67.66	7.66E-10	6.05E-06
telomere capping (GO:0016233)	19	3	0.04	67.66	1.77E-05	1.16E-02
negative regulation of telomere maintenance via telomerase (GO:0032211)	21	3	0.05	61.21	2.32E-05	1.36E-02
negative regulation of telomere maintenance via telomere lengthening (GO:1904357)	28	4	0.07	61.21	8.89E-07	1.56E-03
positive regulation of nitric-oxide synthase activity (GO:0051000)	23	3	0.05	55.89	2.97E-05	1.67E-02
negative regulation of cell aging (GO:0090344)	25	3	0.06	51.42	3.73E-05	1.96E-02
regulation of telomere capping (GO:1904353)	26	3	0.06	49.44	4.15E-05	2.05E-02
regulation of telomere maintenance via telomere lengthening (GO:1904356)	63	7	0.15	47.61	2.55E-10	4.02E-06
positive regulation of telomere maintenance via telomere lengthening (GO:1904358)	37	4	0.09	46.32	2.46E-06	2.59E-03
positive regulation of monooxygenase activity (GO:0032770)	29	3	0.07	44.33	5.61E-05	2.39E-02
telomere maintenance via telomere lengthening (GO:0010833)	30	3	0.07	42.85	6.16E-05	2.49E-02
replication fork processing (GO:0031297)	31	3	0.07	41.47	6.75E-05	2.66E-02
2'-deoxyribonucleotide metabolic process (GO:0009394)	32	3	0.07	40.17	7.37E-05	2.77E-02
regulation of telomere maintenance via telomerase (GO:0032210)	55	5	0.13	38.95	2.75E-07	5.43E-04

Extra rows are shown on the next page

GO biological process complete	Total gene counts	Observed count	expected %	fold Enrichment	*p*-value	FDR
positive regulation of telomere maintenance via telomerase (GO:0032212)	34	3	0.08	37.81	8.72E-05	3.13E-02
deoxyribose phosphate metabolic process (GO:0019692)	34	3	0.08	37.81	8.72E-05	3.06E-02
regulation of telomere maintenance (GO:0032204)	82	7	0.19	36.58	1.41E-09	7.45E-06
RNA-dependent DNA biosynthetic process (GO:0006278)	37	3	0.09	34.74	1.10E-04	3.71E-02
deoxyribonucleotide metabolic process (GO:0009262)	37	3	0.09	34.74	1.10E-04	3.63E-02
positive regulation of telomere maintenance (GO:0032206)	50	4	0.12	34.28	7.53E-06	5.94E-03
negative regulation of DNA biosynthetic process (GO:2000279)	39	3	0.09	32.96	1.28E-04	4.12E-02
regulation of cellular senescence (GO:2000772)	40	3	0.09	32.14	1.37E-04	4.33E-02
DNA-dependent DNA replication maintenance of fidelity (GO:0045005)	41	3	0.1	31.35	1.47E-04	4.46E-02
telomere maintenance (GO:0000723)	99	7	0.23	30.3	4.86E-09	1.92E-05
telomere organization (GO:0032200)	102	7	0.24	29.41	5.91E-09	1.87E-05
protein localization to chromosome (GO:0034502)	64	4	0.15	26.78	1.89E-05	1.20E-02
DNA biosynthetic process (GO:0071897)	109	5	0.25	19.66	6.73E-06	5.91E-03
regulation of DNA biosynthetic process (GO:2000278)	109	5	0.25	19.66	6.73E-06	5.59E-03
negative regulation of chromosome organization (GO:2001251)	137	6	0.32	18.77	9.74E-07	1.54E-03
negative regulation of DNA metabolic process (GO:0051053)	154	6	0.36	16.69	1.88E-06	2.12E-03
positive regulation of DNA metabolic process (GO:0051054)	233	7	0.54	12.87	1.31E-06	1.88E-03
positive regulation of chromosome organization (GO:2001252)	172	5	0.4	12.46	5.62E-05	2.34E-02
regulation of response to DNA damage stimulus (GO:2001020)	214	5	0.5	10.01	1.53E-04	4.56E-02
DNA recombination (GO:0006310)	215	5	0.5	9.96	1.56E-04	4.57E-02
regulation of chromosome organization (GO:0033044)	345	8	0.81	9.94	1.43E-06	1.74E-03
anatomical structure homeostasis (GO:0060249)	337	7	0.79	8.9	1.40E-05	1.00E-02
regulation of DNA metabolic process (GO:0051052)	424	8	0.99	8.08	6.37E-06	5.92E-03

Extra rows are shown on the next page

GO biological process complete	Total gene counts	Observed count	expected %	fold Enrichment	*p*-value	FDR
DNA metabolic process (GO:0006259)	801	13	1.87	6.95	2.85E-08	7.50E-05
chromosome organization (GO:0051276)	1036	11	2.42	4.55	2.24E-05	1.36E-02
organophosphate metabolic process (GO:0019637)	1068	10	2.49	4.01	1.58E-04	4.53E-02
regulation of organelle organization (GO:0033043)	1272	11	2.97	3.71	1.41E-04	4.36E-02
nucleic acid metabolic process (GO:0090304)	2309	16	5.39	2.97	4.41E-05	2.11E-02
nucleobase-containing compound metabolic process (GO:0006139)	2960	18	6.91	2.61	6.99E-05	2.69E-02
heterocycle metabolic process (GO:0046483)	3128	19	7.3	2.6	4.01E-05	2.04E-02
cellular aromatic compound metabolic process (GO:0006725)	3172	19	7.4	2.57	4.88E-05	2.27E-02
organic cyclic compound metabolic process (GO:1901360)	3391	20	7.91	2.53	3.46E-05	1.88E-02
organelle organization (GO:0006996)	3342	19	7.8	2.44	1.01E-04	3.46E-02

Supplementary Table 12: LD score regression (*p*-value<0.05). Genome-wide genetic correlations between LTL and different traits.

Trait	PMID	Study	n	Year	r	se	z	p
Maternal smoking around birth	0	UKBB_Ben_Neale	289727	2017	-0.18	0.06	-3.17	1.50E-03
LDL cholesterol	20686565	GLGC	95454	2010	-0.24	0.08	-2.96	3.10E-03
Diagnoses - main ICD10: E04 Other non-toxic goitre	0	UKBB_Ben_Neale	337199	2017	0.34	0.12	2.89	3.90E-03
Age of first birth	27798627	SSGAC	222037	2016	0.16	0.06	2.83	4.60E-03
Ulcerative colitis	26192919	IIBDGC	27432	2015	0.20	0.07	2.75	0.01
Overweight	23563607	GIANT	158855	2013	-0.14	0.05	-2.71	0.01
Anorexia Nervosa	24514567	GCAN	17767	2014	0.16	0.06	2.68	0.01
HDL cholesterol	20686565	GLGC	100184	2010	0.18	0.07	2.67	0.01
Waist-to-hip ratio	25673412	GIANT	212244	2015	-0.13	0.05	-2.62	0.01
PGC cross-disorder analysis	23453885	PGC	61220	2013	0.16	0.06	2.60	0.01
Infant head circumference	22504419	EGG	10768	2012	0.32	0.13	2.50	0.01
Coronary artery disease	26343387	Cardiogram	184035	2015	-0.14	0.06	-2.49	0.01
Illnesses of father: Heart disease	0	UKBB_Ben_Neale	298237	2017	-0.15	0.06	-2.49	0.01
Urate	23263486	GUGC	110347	2013	-0.11	0.04	-2.44	0.01
Platelet count	22139419	HaemGen	48666	2011	0.16	0.07	2.39	0.02
Smoking status: Previous	0	UKBB_Ben_Neale	336024	2017	-0.12	0.05	-2.34	0.02
Ever smoked	0	UKBB_Ben_Neale	336067	2017	-0.11	0.05	-2.33	0.02
Mothers age at death	0	UKBB_Ben_Neale	199690	2017	0.21	0.09	2.32	0.02
College completion	23722424	SSGAC	95427	2013	0.15	0.07	2.26	0.02
Body mass index	20935630	GIANT	123912	2010	-0.11	0.05	-2.22	0.03
Years of schooling 2016	27225129	SSGAC	293723	2016	0.09	0.04	2.22	0.03
Diagnoses - main ICD10: L03 Cellulitis	0	UKBB_Ben_Neale	337199	2017	-0.35	0.16	-2.18	0.03
Waist circumference	25673412	GIANT	232101	2015	-0.10	0.05	-2.18	0.03
Age at last live birth	0	UKBB_Ben_Neale	123676	2017	0.13	0.06	2.17	0.03
Total Cholesterol	20686565	GLGC	99900	2010	-0.14	0.07	-2.16	0.03
Celiac disease	20190752	NA	15283	2010	-0.22	0.10	-2.15	0.03
Schizophrenia	25056061	PGC	77096	2014	0.09	0.04	2.12	0.03
Cancer code_ self-reported: malignant melanoma	0	UKBB_Ben_Neale	337159	2017	0.31	0.15	2.09	0.04
Smoking status: Current	0	UKBB_Ben_Neale	336024	2017	-0.11	0.05	-2.03	0.04
Diagnoses - main ICD10: J22 Unspecified acute lower respiratory infection	0	UKBB_Ben_Neale	337199	2017	-0.38	0.19	-2.01	0.04
Illnesses of father: None of the above (group 1 - Heart disease, high blood pressure, Chronic bronchitis/emphysema, Alzheimers disease/dementia, Diabetes)	0	UKBB_Ben_Neale	294791	2017	0.19	0.10	1.99	0.05
Weight	0	UKBB_Ben_Neale	336227	2017	-0.07	0.03	-1.99	0.05
Age at Menarche	25231870	ReproGen	182416	2014	0.10	0.05	1.99	0.05
Qualifications: CSEs or equivalent	0	UKBB_Ben_Neale	334070	2017	-0.16	0.08	-1.97	0.05
Whole body fat mass	0	UKBB_Ben_Neale	330762	2017	-0.07	0.03	-1.97	0.05
Hand grip strength (right)	0	UKBB_Ben_Neale	335842	2017	-0.09	0.04	-1.96	0.05
Alanine	27005778	MAGNETIC	24796	2016	0.22	0.11	1.95	0.05
Trunk fat mass	0	UKBB_Ben_Neale	331093	2017	-0.07	0.03	-1.94	0.05
Diagnoses - main ICD10: N20 Calculus of kidney and ureter	0	UKBB_Ben_Neale	337199	2017	0.21	0.11	1.93	0.05

Supplementary Table 13: Case definition for 122 diseases manually curated within UK Biobank.

Disease group	Phenotype	Definition
Cardiovascular diseases	Coronary artery diseases (CAD)	Self-reported history of heart attack/myocardial infarction, coronary angioplasty (PTCA) stent, coronary artery bypass grafts (CABG) or triple heart bypass; or hospitalization for ICD9 410-412, 414, ICD10 I21-I25, OPCS-4 K40-K46, K49, K50.1, K75, or cause of death ICD10 I21-I25
	Atrial fibrillation (AF)	Self-reported history of atrial fibrillation or atrial flutter, or hospitalization or death due to ICD9 427.3, ICD10 I48
	Heart failure (HF)	Self-reported history of heart failure/pulmonary odema, or hospitalization or death due to ICD9 428, ICD10 I50
	Peripheral vascular disease (PVD)	Self-reported history of peripheral vascular disease (PVD) or leg claudication/ intermittent claudication, or hospitalization or death due to ICD9 443.9, 444, ICD10 I73.9,I74
	Venous thromboembolism	Self-reported history of venous thromboembolic disease, pulmonary embolism or deep venous thrombosis (DVT), or hospitalization or death due to ICD9 415.1, 451-453, ICD10 I26,I80-I82
	Aortic valve stenosis	Self-reported history of aortic stenosis, or hospitalization or death due to ICD9 424.1, ICD10 I35.0
	Hypertensive diseases	Self-reported use of blood pressure medications, or systolic blood pressure >140 mmHg or diastolic blood pressure >90 mmHg, or hospitalization or death due to ICD9 401-405, ICD10 I10-I13,I15
	Stroke	Self-reported history of stroke, subarachnoid haemorrhage or ischaemic stroke, or hospitalization or death due to ICD9 430-432, ICD10 I60-I64
	Varicose veins	Self-reported history of varicose veins or varicose ulcer, hospitalization or deat due to ICD-10: I83, I84, ICD-9-CM: 454, OPER code 1479 varicose vein surgery
	Raynaud's phenomenon/disease	Self-reported history of raynaud's phenomenon/disease, ICD-10: I73.0 ICD-9-CM: 443.0
Endocrine disorders	Diabetes	Self-reported diabetes, type 1 or type 2 diabetes or hospitalization or death due to ICD9 250, ICD10 E10-E11,E13-E14
	Diabetes type I	Self-reported type 1 diabetes or hospitalization or death due to ICD9 250 (juvenile type - 250.01, 250.03, 250.11, 250.13, 250.21, 250.23, 250.31, 250.33, 250.41, 250.43, 250.51, 250.53, 250.61, 250.63, 250.71, 250.73, 250.81, 250.83, 250.91, 250.93), ICD10 E10
	Diabetes type II	Self-reported generic or type 2 diabetes and age of onset 35+ years old, or hospitalization or death due to ICD9 250 (non juvenile type), ICD10 E11,E13-E14
	Hyperthyroid	Self-reported history of hyperthyroidism/thyrotoxicosis or hospitalization or death due to ICD9 242.9, ICD10 E05
	Hypothyroid	Self-reported history of hypothyroidism/myxoedema or hospitalization or death due to ICD9 244.9, ICD10 E03.9

Extra rows are shown on the next page

Mental illnesses	Anxiety	Self-reported history of anxiety/panic attacks or hospitalization or death due to ICD9 300.0, ICD10 F41
	Depression	Self-reported history of depression or hospitalization or death due to ICD9 296.2-296.3, ICD10 F32-F33
	Multiple sclerosis	Self-reported history of multiple sclerosis or hospitalization or death due to ICD9 340, ICD10 G35
	Epilepsy	Self-reported history of epilepsy or hospitalization or death due to ICD9 345, ICD10 G40-G41
	Dementia	Self-reported history of dementia/alzheimers/cognitive impairment, or hospitalization or death due to ICD9 290,330-331, ICD10 F00-F03,G30-G31
	Parkinsons' disease	Self-reported history of Parkinson's disease, or hospitalization or death due to ICD9 332, ICD10 G20-G21
	Migraine	Self-reported history of migraine, or hospitalization due to ICD9 346, ICD10 G43
	Mania/bipolar disorder/manic depression	Self-reported history of mania/bipolar disorder/manic depression, ICD10 F30-F31, ICD9 296.0-296.1, 296.4-296.8
	Anorexia nervosa	Self-reported history of anorexia, ICD-10: F500,F502,F508,R630, ICD-9-CM: 3071,7830,30751
	Schizophrenia	Self-reported history of schizophrenia, ICD-10: F20, ICD-9-CM: 295
	Chronic fatigue syndrome	Self-reported history of chronic fatigue syndrome, ICD10 R5382, ICD9 78071
Digestive diseases	Gastro-oesophageal reflux disease (GORD)	Self-reported history of gastro-oesophageal reflux or gastric reflux, or hospitalization or death due to ICD9 530.11, 530.81, ICD10 K21
	Irritable bowel syndrome (IBS)	Self-reported history of irritable bowel syndrome, or hospitalization or death due to ICD9 564.1, ICD10 K58
	Inflammatory bowel disease (IBD)	Self-reported history of inflammatory bowel disease, Crohn's disease, or ulcerative colitis, or hospitalization or death due to ICD9 555-556, ICD10 K50-K51
	Gallstone	Self-reported history of cholelithiasis/gall stones, or hospitalization or death due to ICD9 574, ICD10 K80, or OPER4 code 1455 cholecystectomy/gall bladder removal, 1528 gallstones removed
	Peptic ulcer	Self-reported history of peptic ulcer, duodenal ulcer or gastric/stomach ulcers, or hospitalization or death due to ICD9 531-533, ICD10 K25-K27, OPER code 1566 peptic ulcer surgery, 1567 gastric ulcer surgery
	Liver cirrhosis	Self-reported history of liver failure/cirrhosis, primary biliary cirrhosis, alcoholic liver disease or alcoholic cirrhosis, or hospitalization or death due to ICD9 571, ICD10 K70, K74
	Appendicitis	Self-reported history of appendicitis, or hospitalization or death due to ICD9 540-543, ICD10 K35-K37
	Oesophagitis/barretts oesophagus	Self-reported history of oesophagitis/barretts oesophagus, ICD10 K20 and ICD9 530.10 (oesophagitis), ICD-10: K22.7 and ICD-9-CM: 530.85 (barretts oesophagus)
	Hiatus hernia	Self-reported history of hiatus hernia, ICD-10: K44.0,K44.1,K44.9, ICD-9-CM: 552.3, 553.3, 551.3
	Abdominal hernia	Self-reported history of abdominal hernia, ICD-10: K45-K46
	Umbilical hernia	Self-reported history of umbilical hernia, ICD-10: K42, ICD-9-CM: 5511, 5521, 5531
	Inguinal hernia	Self-reported history of inguinal hernia, ICD-10: K40, ICD-9-CM: 5500,5501,5509
	Malabsorption/coeliac disease	Self-reported history of coeliac disease, ICD-10: K90.0 ICD-9-CM: 579.0
	Diverticular disease/diverticulitis	Self-reported history of diverticular disease/diverticulitis, ICD-10: K57 ICD-9-CM: 562
	Rectal or colon adenoma/polyps	Self-reported history of rectal or colon adenoma/polyps or benign neoplasms, ICD10 K63.5,K62.1,D12, ICD9 5690,211.3,211.4,2095, OPCS4 H481 - Excision of polyp of anus
	Haemorrhoids / piles	Self-reported history of haemorroids, OPER code 1483 haemorroidectomy / piles surgery/ banding of piles, ICD-10: K64 ICD-9-CM: 455
	Pancreatitis	Self-reported history of pancreatitis, ICD-10: K85, K86.0–K86.1, B25.2,B26.3,K87.1, ICD-9-CM: 577.0–577.1, 0723
	Peritonitis	Self-reported history of peritonitis, ICD-10: K65,K67,N733,N734,N735,A1831,A5485,A7481, ICD-9-CM: 567,56889,0140,03283,0952,09886,6145,6147

Genito-urinary diseases	Chronic kidney diseases	Self-reported history of renal/kidney failure requiring or not requiring dialysis, or hospitalization or death due to ICD9 585, ICD10 N18
	Benign prostatic hyperplasia (BPH)	(Male only) Self-reported history of enlarged prostate or benign prostatic hypertrophy (BPH), or hospitalization or death due to ICD9 600, ICD10 N40
	Uterine fibroid	(Female only) Self-reported history of uterine fibroids, or hospitalization or death due to ICD9 218, ICD10 D25, 1509 myomectomy/fibroids removed
	Kidney stone/ureter stone/bladder stone	Self-reported history of kidney stone/ureter stone/bladder stone, OPER code 1197 percutaneous/open kidney stone surgery/lithotripsy, ICD-10: N20.0 – N20.9,N21,N22,N13.2, ICD-9-CM: 592.0, 592.1,592.9,594
	Female infertility	Female-only. Self-reported history of female infertility, ICD-10 N97.0, ICD-9-CM 628
	Ovarian cyst	Female-only. Self-reported history of 1349 ovarian cyst or cysts, 1350 polycystic ovaries/polycystic ovarian syndrome, OPER code 1506 ovarian cyst removal/surgery, OPCS Q474 (open drainage of cyst of ovary) and Q493 (endoscopic drainage of cyst of ovary), ICD-10: N83.0-N83.2,E282,D27, ICD-9-CM: 620.0-620.2,2564,
	Uterine polyps	Female-only. Self-reported history of uterine polyps, ICD10 N84.0,N84.1,D26, ICD9 6210,2190,2191, OPER code 1539 uterine polypectomy/uterine polyps removed
	Vaginal prolapse/uterine prolapse	Female-only. Self-reported history of vaginal prolapse/uterine prolapse, ICD10 N81, ICD9 618.0-618.4,618.6-618.9
	Endometriosis	Female-only. Self-reported history of endometriosis, ICD-10: N80 ICD-9-CM: 617
	Breast cyst	Female only. Self-reported history of breast cysts, ICD10 N60.0-N60.4, ICD9 610.0-610.4, OPER code 1513 breast cyst/abscess removal
	Benign breast lump	Female only. Self-reported history of breast lump, ICD10 D24,N608,N609, ICD9 217,6108,6109
Musculoskeletal diseases	Gout	Self-reported history of gout, or hospitalization or death due to ICD9 274, ICD10 M10
	Rheumatoid arthritis	Self-reported history of rheumatoid arthritis, or hospitalization or death due to ICD9 714, ICD10 M05-M06
	Osteoarthritis	Self-reported history of osteoarthritis, or hospitalization or death due to ICD9 715, ICD10 M15-M19
	Osteoporosis	Self-reported history of osteoporosis, or hospitalization or death due to ICD9 733.0, ICD10 M80-M82
	Sciatica	Self-reported history of sciatica, or hospitalization or death due to ICD9 724.3, ICD10 M54.3-M54.4
	Intervertebral disc disorder - prolased disc / degenerative disc	Self-reported history of prolapsed disc/slipped disc or disc degeneration, or hospitalization or death due to ICD9 722, ICD10 M50-M51
	Spine arthritis/spondylitis	Self-reported history of spine arthritis/spondylitis (ICD-10 M46.0,M46.1,M46.5-M46.9, ICD-9-CM 721.90,721.91) or ankylosing spondylitis (ICD-10: M08.1, M45, ICD-9-CM: 720.0)

Extra rows are shown on the next page

Respiratory diseases	Chronic obstructive pulmonary disease (COPD)	Self-reported history of COPD,emphysema/chronic bronchitis, or hospitalization or death due to ICD9 490-492,495-496, ICD10 J40-J44
	Asthma	Self-reported history of asthma, or hospitalization or death due to ICD9 493, ICD10 J45-J46
	Lower respiratory infection / pneumonia	Self-reported history of pneumonia, or hospitalization or death due to ICD9 466,480-487, ICD10 J10-J18,J20-J22
	Otitis media	Self-reported history of otitis media, or hospitalization or death due to ICD9 381-382, ICD10 H65-H66
	Hayfever_eczema	Self-reported history of hayfever, allergic rhinitis, eczema or contact dermatitis
	Bronchiectasis	Self-reported history of bronchiectasis, ICD-10: J47, Q33.4; ICD-9-CM: 494, 748.61
	Sleep apnoea	Self-reported history of sleep apnoea, ICD-10: G47.3, ICD-9-CM: 327.2, 780.57
	Pleurisy	Self-reported history of pleurisy, hospitalisation or death due to ICD-10: R09.1, ICD-9-CM: 511.0,511.1
	Pneumothorax	Self-reported history of spontaneous pneumothorax/recurrent pneumothorax, hospitalisation or death due to ICD-10: J93.0,J93.1,J9381 , ICD-9-CM: 512.0,512.81,512.82
	Chronic sinusitis	Self-reported history of chronic sinusitis, ICD-10: J01, J32, ICD-9-CM: 461, 473
	Nasal polyps	Self-reported history of nasal polyps, hospitalisation due to ICD10 J33, or ICD9 471, OPER codes 1559 nasal polyp surgery / nasal polypectomy
	Tonsillitis	Self-reported history of tonsillitis, hospitalisation due to ICD-10: J03, J35.0 ICD-9-CM: 463, 474.0, OPER code 1478 tonsillectomy +/- adenoids
	Meniere's disease	Self-reported history of meniere's disease, ICD-10: H81.0 ICD-9-CM: 386.0
	Tinnitus	Self-reported history of tinnitus, ICD-10: H93.1 ICD-9-CM: 388.3
Infections and others	Rheumatic fever	Self-reported rheumatic fever, ICD10 I00-I02, ICD9 390 (rheum fever) and 391 (with heart involvement), rheumatic chorea ICD9 392
	Meningitis	Self-reported history of meningitis, ICD-10: G00–G03 ICD-9-CM: 320–321
	Measles / morbillivirus	Self-reported history of measles / morbillivirus, ICD-10: B05 ICD-9-CM: 055
	Rubella / german measles	Self-reported history of rubella / german measles, ICD-10: B06 ICD-9-CM: 056
	Chickenpox	Self-reported history of chickenpox, ICD-10: B01 ICD-9-CM: 052
	Shingles	Self-reported history of shingles, ICD-10: B02 ICD-9-CM: 053
	Infectious mononucleosis / glandular fever / epstein barr virus (ebv)	Self-reported history of infectious mononucleosis / glandular fever / epstein barr virus (ebv), CD-10: B27 ICD-9-CM: 075
	Mumps / epidemic parotitis	Self-reported history of mumps / epidemic parotitis, ICD-10: B26 ICD-9-CM: 072
	Helicobacter pylori	Self-reported history of helicobacter pylori, ICD-10: B9681, ICD-9-CM: 041.86
	Tuberculosis (tb)	Self-reported history of tuberculosis (tb), ICD-10: A15–A19 ICD-9-CM: 010–018
	Whooping cough / pertussis	Self-reported history of whooping cough / pertussis, ICD-10: A37 ICD-9-CM: 033
	Scarlet fever / scarlatina	Self-reported history of scarlet fever / scarlatina, ICD-10: A38 ICD-9-CM: 034.1
	Malaria	Self-reported history of malaria, ICD-10: B50-B54 ICD-9-CM: 084
Eye Problems	Retinal detachment	Self-reported history of retinal detachment, ICD-10: H330,H332,H334, ICD-9-CM: 3610,3612,3618
	Diabetic eye disease	Self-reported history of diabetic eye disease, ICD-10 H36 (E10.3 E11.3 E12.3 E13.3 E14.3), ICD-9-CM 250.5, 3620,36641
	Glaucoma	Self-reported history of glaucoma, ICD-10: H40-H42 ICD-9-CM: 365, OPER code 1436 glaucoma surgery/trabeculectomy
	Cataract	Self-reported history of cataract, ICD-10: H25-H26, H28, Q12.0 ICD-9-CM: 366, OPER code 1435 cataract extraction/lens implant

Extra rows are shown on the next page

Immune / inflammatory	Sarcoidosis	Self-reported history of sarcoidosis, ICD-10: D86 ICD-9-CM: 135
	Psoriasis	Self-reported history of psoriasis, ICD-10: L40 ICD-9-CM: 696
	Allergy/hypersensitivity/anaphylaxis	Self-reported history of allergy/hypersensitivity/anaphylaxis (combines all allergies and anahylaxis) General: ICD-10: T78.2, T78.4, ICD-9-CM: 995.0, V1381 (anaphylaxis) To food: ICD10 T780-T781, Z9101, Z9102, ICD9 9956, 997 (please see ICD codes description in following phenotype) To drugs: ICD10 T886, Z88, ICD9 99527 (please see ICD codes description in following phenotype) Additional: ICD10 Z91103-Z9109, K0855, ICD9 52566, 9953, V150
	Allergy or anaphylactic reaction to food	Self-reported history of allergy or anaphylactic reaction to food, ICD10 T780-T781, Z9101, Z9102, ICD9 9956, 997
	Allergy or anaphylactic reaction to drug	Self-reported history of allergy or anaphylactic reaction to drug, ICD10 T886, T887, Z88, ICD9 99527
	Polymyalgia rheumatica	Self-reported history of polymyalgia rheumatica, ICD-10 M35.3, ICD-9-CM 725
	Systemic lupus erythematosis/sle	Self-reported history of systemic lupus erythematosis/sle, ICD-10: M32,H0112, L93, ICD-9-CM: 710.0,37334,6954
	Sjogren's syndrome/sicca syndrome	Self-reported history of sjogren's syndrome/sicca syndrome, ICD-10: M35.0 ICD-9-CM: 710.2

Extra rows are shown on the next page

Cancer	Lung cancer	Self-reported history of lung cancer, small cell or non-small cell lung cancer or trachea cancer, or cancer registration or death due to ICD9 162, ICD10 C33-C34
	Colorectal cancer	Self-reported history of large bowel/colorectal cancer, colon cancer/sigmoid cancer, rectal cancer or anal cancer, or cancer registration or death due to ICD9 153, 154.0-154.1, ICD10 C18-C20
	Breast cancer	(Female only) Self-reported history of breast cancer, or cancer registration or death due to ICD9 174, ICD10 C50
	Prostate cancer	(Male only) Self-reported history of prostate cancer, or cancer registration or death due to ICD9 185, ICD10 C61
	Thyroid cancer	Self-reported history of thyroid cancer, or cancer registration or death due to ICD9 193, ICD10 C73
	Oesophageal cancer	Self-reported history of oesophageal cancer, or cancer registration or death due to ICD9 150, ICD10 C15
	Stomach cancer	Self-reported history of stomach cancer, or cancer registration or death due to ICD9 151, ICD10 C16
	Pancreas cancer	Self-reported history of pancreas cancer, or cancer registration or death due to ICD9 157, ICD10 C25
	Melanoma	Self-reported history of malignant melanoma, or cancer registration or death due to ICD9 172, ICD10 C43, OPER code 1593 removal of malignant melanoma
	Skin cancer (including melanoma)	Self-reported history of skin cancer, malignant melanoma, non-melanoma skin cancer, basal cell carcinoma or squamous cell carcinoma, or cancer registration or death due to ICD9 172-173, ICD10 C43-C44, OPER codes 1595 removal of squamous cell carcinoma (scc), 1593 removal of malignant melanoma, 1596 removal of rodent ulcer / basal cell carcinoma (bcc)
	Cervical cancer	Self-reported history of cervical cancer, or cancer registration or death due to ICD9 180, ICD10 C53
	Uterus cancer	Self-reported history of uterine/endometrial cancer, or cancer registration or death due to ICD9 179,182, ICD10 C54-C55
	Ovary cancer	Self-reported history of ovarian cancer or fallopian tube cancer, or cancer registration or death due to ICD9 183, ICD10 C56-C57.4
	Kidney cancer	Self-reported history of kidney/renal cell cancer, or cancer registration or death due to ICD9 189, ICD10 C64-C66,C68
	Bladder cancer	Self-reported history of bladder cancer, or cancer registration or death due to ICD9 188, ICD10 C67
	Non-Hodgkin lymphoma	Self-reported history of non-Hodgkins lymphoma, or cancer registration or death due to ICD9 200,202, ICD10 C82-C86
	Lymphomas and multiple myeloma	Self-reported history of lymphoma, Hodgkins or non-Hodgkins lymphoma, multiple myeloma, or cancer registration or death due to ICD9 200-203, ICD10 C81-C88,C90,C96
	Leukaemia	Self-reported history of leukaemia, acute myeloid leukaemia, chronic lymphocytic or chronic myeloid, or cancer registration or death due to ICD9 204-208, ICD10 C91-C95
	Brain cancer / primary malignant brain tumour	Self-reported history of brain cancer / primary malignant brain tumour, or cancer registration or death due to ICD-10: C71, ICD-9-CM: 191
	Head and neck cancer	Self-reported history of larynx/throat cancer, parotid gland cancer, other salivary gland cancer, lip cancer, tongue cancer, gum cancer, mouth cancer, tonsil cancer, oropharynx / oropharyngeal cancer, nasal cavity cancer, sinus cancer, or cancer registration or death due to ICD-10 C32, ICD-9-CM 161 (laryngeal cancer), C30 (nasal cavity), and C00–C14/D10–D11, 140–149/210 (head and neck cancers)
	Testicular cancer	Self-reported history of testicular cancer, or cancer registration or death due to ICD-10 C62, ICD-9-CM 186

Supplementary Table 14: Estimated power to detect an odds ratio (OR) in the range of 0.9 to 1.1 for given numbers of cases within UK Biobank.

Phenotype	N(Cases)	Detectable OR							
		0.9	0.95	0.975	0.99	1.01	1.025	1.05	1.1
abdominal hernia	753	82%	29%	10%	5%	5%	10%	27%	74%
allergy hypersensitivity	32,232	100%	100%	99%	41%	41%	99%	100%	100%
allergy to drug	31,852	100%	100%	99%	41%	40%	99%	100%	100%
allergy to food	2,333	100%	70%	23%	7%	7%	22%	65%	100%
anorexia	1,264	96%	44%	14%	5%	5%	14%	41%	92%
anxiety	12,253	100%	100%	79%	19%	19%	77%	100%	100%
aortic valve stenosis	1,894	100%	61%	19%	6%	6%	19%	56%	99%
appendicitis	7,840	100%	99%	60%	14%	14%	58%	99%	100%
asthma	59,305	100%	100%	100%	63%	62%	100%	100%	100%
atrial fibrillation	16,439	100%	100%	89%	24%	24%	87%	100%	100%
benign breast lump	3,786	100%	88%	34%	9%	9%	33%	85%	100%
benign prostatic hyperplasia	16,562	100%	100%	89%	25%	24%	88%	100%	100%
bladder cancer	2,529	100%	73%	25%	7%	7%	24%	69%	100%
brain cancer	688	79%	27%	10%	4%	4%	9%	25%	70%
breast cancer	15,018	100%	100%	86%	23%	22%	85%	100%	100%
breast cyst	6,906	100%	99%	55%	13%	13%	53%	98%	100%
bronchiectasis	2,564	100%	74%	25%	7%	7%	24%	69%	100%
cad	31,486	100%	100%	99%	41%	40%	99%	100%	100%
cataract	27,998	100%	100%	98%	37%	36%	98%	100%	100%
cervical cancer	4,570	100%	93%	40%	10%	10%	38%	91%	100%
chickenpox	3,196	100%	82%	30%	8%	8%	28%	78%	100%
chronic fatigue syndrome	2,012	100%	63%	20%	7%	6%	20%	59%	99%
chronic kidney disease	5,536	100%	97%	47%	11%	11%	45%	95%	100%
coeliac disease	2,713	100%	76%	26%	8%	7%	25%	72%	100%
colorectal cancer	5,558	100%	97%	47%	11%	11%	45%	95%	100%
colorectal polyp	25,760	100%	100%	98%	35%	34%	97%	100%	100%
copd	15,032	100%	100%	86%	23%	22%	85%	100%	100%
dementia	1,681	99%	56%	18%	6%	6%	17%	51%	97%
depression	34,400	100%	100%	99%	43%	43%	99%	100%	100%
diabetes	30,804	100%	100%	99%	40%	39%	99%	100%	100%
diabetes1	3,469	100%	85%	32%	9%	8%	30%	82%	100%
diabetes2	20,576	100%	100%	94%	29%	29%	93%	100%	100%
diabetic eye disease	2,643	100%	75%	25%	7%	7%	24%	71%	100%
diverticulitis	31,164	100%	100%	99%	40%	40%	99%	100%	100%
endometriosis	7,312	100%	99%	57%	13%	13%	55%	99%	100%
epilepsy	5,560	100%	97%	47%	11%	11%	45%	95%	100%
female infertility	1,003	92%	37%	12%	5%	5%	12%	34%	85%
gallstone	26,233	100%	100%	98%	35%	35%	97%	100%	100%
gastro gord	40,496	100%	100%	100%	49%	48%	100%	100%	100%
glaucoma	8,143	100%	100%	62%	14%	14%	60%	99%	100%
gout	8,373	100%	100%	63%	15%	15%	61%	99%	100%

Phenotype	N(Cases)	Detectable OR							
		0.9	0.95	0.975	0.99	1.01	1.025	1.05	1.1
haemorrhoids	9,132	100%	100%	67%	16%	15%	65%	100%	100%
hayfever eczema	113,143	100%	100%	100%	83%	83%	100%	100%	100%
head and neck cancer	2,636	100%	75%	25%	7%	7%	24%	70%	100%
heart failure	6,113	100%	98%	50%	12%	12%	48%	97%	100%
helicobacter pylori	1,334	97%	46%	15%	6%	6%	14%	43%	94%
hiatus hernia	33,483	100%	100%	99%	42%	42%	99%	100%	100%
hypertension	232,147	100%	100%	100%	92%	92%	100%	100%	100%
hyperthyroid	5,023	100%	95%	43%	11%	10%	41%	93%	100%
hypothyroid	26,729	100%	100%	98%	36%	35%	97%	100%	100%
inflamatory bowel disease	6,163	100%	98%	51%	12%	12%	49%	97%	100%
inguinal hernia	18,516	100%	100%	92%	27%	26%	91%	100%	100%
intervertebral disc disorder	16,648	100%	100%	89%	25%	24%	88%	100%	100%
irritable bowel syndrome	15,175	100%	100%	87%	23%	23%	85%	100%	100%
kidney cancer	1,412	98%	49%	16%	6%	6%	15%	45%	95%
kidney stone	9,168	100%	100%	67%	16%	15%	65%	100%	100%
leukemia	1,306	97%	46%	15%	6%	5%	14%	42%	93%
liver cirrhosis	2,614	100%	74%	25%	7%	7%	24%	70%	100%
lung cancer	2,485	100%	72%	24%	7%	7%	23%	68%	100%
lupus	812	85%	31%	11%	5%	5%	10%	28%	77%
lymphomas	3,300	100%	84%	30%	8%	8%	29%	80%	100%
malaria	779	84%	30%	10%	5%	5%	10%	27%	76%
mania	1,882	100%	60%	19%	6%	6%	19%	56%	98%
measles	2,659	100%	75%	26%	7%	7%	24%	71%	100%
melanoma	4,766	100%	94%	41%	10%	10%	40%	92%	100%
meniere disease	1,554	99%	52%	17%	6%	6%	16%	48%	96%
meningitis	2,050	100%	64%	21%	7%	7%	20%	60%	99%
migraine	16,022	100%	100%	88%	24%	23%	87%	100%	100%
mononucleosis	757	83%	29%	10%	5%	5%	10%	27%	75%
multiple sclerosis	1,951	100%	62%	20%	6%	6%	19%	58%	99%
mumps	1,567	99%	53%	17%	6%	6%	16%	49%	96%
nasal polyp	5,526	100%	97%	46%	11%	11%	45%	95%	100%
non.hodgkin lymphoma	2,200	100%	67%	22%	7%	7%	21%	63%	99%
oesophageal cancer	795	84%	30%	11%	5%	5%	10%	28%	77%
oesophagitis	3,623	100%	87%	33%	9%	9%	32%	83%	100%
osteoarthritis	71,185	100%	100%	100%	69%	68%	100%	100%	100%
osteoporosis	12,562	100%	100%	80%	20%	19%	78%	100%	100%
otitis media	1,963	100%	62%	20%	6%	6%	19%	58%	99%
ovarian cyst	13,177	100%	100%	82%	21%	20%	80%	100%	100%
ovary cancer	1,462	98%	50%	16%	6%	6%	15%	46%	95%
pancreatitis	2,816	100%	77%	27%	8%	8%	26%	73%	100%
parkinsons	1,489	98%	51%	16%	6%	6%	16%	47%	96%
peptic ulcer	12,671	100%	100%	80%	20%	20%	78%	100%	100%
periph vascular disease	4,287	100%	92%	38%	10%	9%	36%	89%	100%
peritonitis	2,446	100%	72%	24%	7%	7%	23%	67%	100%
pertussis	757	83%	29%	10%	5%	5%	10%	27%	75%

Phenotype	N(Cases)	Detectable OR							
		0.9	0.95	0.975	0.99	1.01	1.025	1.05	1.1
pleurisy	2,131	100%	66%	21%	7%	7%	21%	61%	99%
pneumonia	23,222	100%	100%	96%	32%	31%	96%	100%	100%
pneumothorax	1,209	96%	43%	14%	5%	5%	14%	40%	91%
polymyalgia rheumatica	1,826	99%	59%	19%	6%	6%	18%	55%	98%
prostate cancer	7,173	100%	99%	57%	13%	13%	55%	98%	100%
psoriasis	6,695	100%	99%	54%	13%	12%	52%	98%	100%
raynauds	4,096	100%	90%	36%	9%	9%	35%	87%	100%
retinal detachment	3,405	100%	85%	31%	8%	8%	30%	81%	100%
rheumatic fever	1,359	97%	47%	15%	6%	6%	15%	43%	94%
rheumatoid arthritis	7,714	100%	99%	60%	14%	14%	58%	99%	100%
rubella	807	85%	31%	11%	5%	5%	10%	28%	77%
sarcoidosis	1,226	96%	43%	14%	5%	5%	14%	40%	92%
scarlatina	684	79%	27%	10%	4%	4%	9%	25%	70%
schizophrenia	957	90%	35%	12%	5%	5%	12%	33%	84%
sciatica	7,018	100%	99%	56%	13%	13%	54%	98%	100%
shingles	1,042	92%	38%	13%	5%	5%	12%	35%	87%
sinusitis	6,201	100%	98%	51%	12%	12%	49%	97%	100%
sjogren	794	84%	30%	11%	5%	5%	10%	28%	77%
skin cancer	19,457	100%	100%	93%	28%	27%	92%	100%	100%
sleep apnoea	5,729	100%	97%	48%	11%	11%	46%	96%	100%
spondilitis	6,449	100%	98%	52%	12%	12%	50%	97%	100%
stomach cancer	710	80%	28%	10%	5%	5%	10%	25%	72%
stroke	11,650	100%	100%	77%	19%	18%	75%	100%	100%
testicular cancer	854	87%	32%	11%	5%	5%	11%	30%	79%
thyroid cancer	671	78%	26%	10%	4%	4%	9%	24%	69%
tinnitus	1,884	100%	60%	19%	6%	6%	19%	56%	98%
tonsilitis	71,654	100%	100%	100%	70%	69%	100%	100%	100%
tuberculosis	2,535	100%	73%	25%	7%	7%	24%	69%	100%
umbilical hernia	4,521	100%	93%	40%	10%	10%	38%	90%	100%
uterine fibroid	19,278	100%	100%	93%	28%	27%	92%	100%	100%
uterine polyps	13,014	100%	100%	81%	20%	20%	79%	100%	100%
uterine prolapse	13,789	100%	100%	83%	21%	21%	81%	100%	100%
uterus cancer	1,993	100%	63%	20%	7%	6%	19%	58%	99%
varicose	48,825	100%	100%	100%	56%	55%	100%	100%	100%
venous thromboembolism	16,244	100%	100%	89%	24%	24%	87%	100%	100%
vertigo	7,873	100%	99%	61%	14%	14%	58%	99%	100%

Supplementary Table 15: Significant associations (*p*-value < 0.05) between genetically predicted LTL and diseases among 122 diseases manually curated in UK Biobank. Nominally significant associations were highlighted in yellow, among which those passed the Bonferroni corrected significance threshold were in red.

Disease group	Phenotype	N cases	IVW	Eggers	Median-MR	RAPS	IVW	Eggers	Median-MR	RAPS	Eggers Intercept p-value
Cardiovascular diseases	Coronary artery diseases (CAD)	31,486	0.01	0.27	0.03	0.02	1.11	1.12	1.14	1.10	0.95
	Hypertensive diseases	232,147	0.02	0.06	1.00E-06	0.01	0.92	0.84	0.85	0.91	0.27
	Aortic valve stenosis	1,894	0.03	0.32	0.13	0.03	1.37	1.39	1.38	1.37	0.96
	Venous thromboembolism	16,244	0.04	0.85	0.04	0.03	0.90	0.98	0.87	0.89	0.45
	Heart failure (HF)	6,113	0.12	0.56	0.02	0.08	1.13	1.12	1.31	1.16	0.93
Endocrine disorders	Hypothyroid	26,729	2.91E-05	0.03	1.20E-05	0.00	1.37	1.49	1.35	1.24	0.60
	Hyperthyroid	5,023	0.01	0.14	0.00	0.03	1.44	1.68	1.55	1.34	0.62
Mental illnesses	Multiple sclerosis	1,951	0.03	0.08	0.04	0.04	0.67	0.46	0.64	0.70	0.36
	Anxiety	12,253	0.04	0.95	0.82	0.07	0.88	0.99	0.98	0.88	0.39
	Chronic fatigue syndrome	2,012	0.12	0.02	0.12	0.11	0.79	0.42	0.71	0.77	0.05
Digestive diseases	Diverticular disease/Diverticulitis	31,164	0.03	0.05	0.02	0.06	0.91	0.81	0.88	0.92	0.22
	Rectal or colon adenoma/polyps	25,760	0.05	0.07	0.01	0.00	0.88	0.76	0.84	0.87	0.27
	Haemorrhoids / piles	9,132	0.08	0.02	0.06	0.05	0.89	0.69	0.83	0.88	0.07
	Malabsorption/Coeliac disease	2,713	0.25	0.08	0.00	0.09	1.86	9.63	1.74	1.55	0.16
	Oesophagitis/barretts oesophagus	3,623	0.31	0.58	0.01	0.04	0.88	1.17	0.65	0.80	0.28
Musculoskeletal diseases	Rheumatoid arthritis	7,714	0.00	0.23	0.03	0.03	1.33	1.34	1.27	1.23	0.98
Respiratory diseases	Bronchiectasis	2,564	0.01	0.13	0.07	0.03	1.35	1.55	1.38	1.31	0.59
	Chronic obstructive pulmonary disease (COPD)	15,032	0.03	0.13	0.19	0.05	1.14	1.25	1.12	1.13	0.49
	Hayfever_eczema	113,143	0.21	0.30	0.02	0.41	1.04	1.09	1.10	1.03	0.56
Infections	Tuberculosis (tb)	2,535	0.01	0.62	0.01	0.03	1.35	1.16	1.70	1.35	0.55
	Meningitis	2,050	0.06	0.74	0.48	0.03	0.76	0.89	0.86	0.73	0.61
Eye Problems	Retinal detachment	3,405	0.17	0.00	0.01	0.13	0.84	0.40	0.65	0.81	0.01
Immune / inflammatory	Sarcoidosis	1,226	0.02	0.13	0.25	0.06	1.53	1.97	1.33	1.40	0.53

Disease group	Phenotype	N cases	IVW	Eggers	Median-MR	RAPS	IVW	Eggers	Median-MR	RAPS	Eggers Intercept p-value
Cancer	Lung cancer	2,485	1.03E-05	0.02	2.66E-04	1.99E-05	0.55	0.47	0.49	0.54	0.60
	Skin cancer (including melanoma)	19,457	1.77E-05	0.00	4.20E-05	5.12E-11	0.69	0.51	0.70	0.72	0.09
	Thyroid cancer	671	4.76E-05	0.04	0.00	3.09E-06	0.35	0.27	0.28	0.31	0.66
	Lymphomas and multiple myeloma	3,300	1.52E-04	0.02	1.92E-05	4.76E-07	0.60	0.45	0.49	0.56	0.32
	Leukaemia	1,306	1.74E-04	0.00	7.92E-05	2.77E-04	0.50	0.23	0.35	0.49	0.05
	Kidney cancer	1,412	0.01	0.00	0.02	0.00	0.57	0.20	0.52	0.51	0.02
	Melanoma	4,766	0.01	0.07	0.00	3.50E-04	0.70	0.54	0.63	0.71	0.39
	Brain cancer / primary malignant brain tumour	688	0.01	0.19	0.18	0.30	0.43	0.34	0.58	0.70	0.76
	Non-Hodgkin lymphoma	2,200	0.01	0.13	0.03	0.00	0.68	0.56	0.63	0.66	0.58
	Bladder cancer	2,529	0.05	0.07	0.00	0.22	0.76	0.54	0.57	0.83	0.27
	Pancreas cancer	676	0.50	0.02	0.37	0.35	1.21	4.58	1.39	1.32	0.02
Female-only	Uterine fibroid	19,273	6.26E-07	1.81E-04	5.70E-05	9.37E-05	0.60	0.39	0.67	0.70	0.04
	Uterine polyps	13,007	8.76E-05	0.02	3.72E-07	5.73E-11	0.73	0.62	0.63	0.68	0.37
	Ovarian cyst	13,161	0.02	0.03	0.13	0.14	0.84	0.68	0.86	0.90	0.17
	Breast cyst	6,747	0.02	0.17	0.06	0.01	0.82	0.75	0.80	0.82	0.64
	Endometriosis	7,312	0.07	0.01	0.01	0.08	0.87	0.63	0.75	0.86	0.05
	Female infertility	1,003	0.13	0.75	0.50	0.15	1.39	1.18	1.24	1.43	0.74
	Cervical cancer	3,250	0.14	0.14	0.04	0.11	1.14	1.38	1.34	1.18	0.34
	Benign breast lump	3,740	0.19	0.37	0.04	0.08	0.87	0.80	0.74	0.83	0.70
Male-only	Benign prostatic hyperplasia (BPH)	16,557	2.09E-04	1.15E-03	2.43E-04	0.01	0.69	0.44	0.68	0.77	0.05
	Prostate cancer	7,168	1.93E-03	0.24	2.96E-04	4.99E-04	0.65	0.68	0.60	0.63	0.90
	Testicular cancer	854	0.22	0.01	0.97	0.36	1.35	4.45	0.99	1.27	0.02

Supplementary Table 16: Definition of 27 cancers based on self-reported disease histories and ICD-10 codes in UK Biobank.

Disease_ID	N_total	N_case	Disease_name	Disease_definition
CAN-0001	352070	3222	Lung/bronchus/trachea cancer	Self-reported history of lung cancer, small cell lung cancer, non-small cell lung cancer, or trachea cancer, or cancer registration or death due to ICD10 C33-C34, C78.0, D02.2, D38.1
CAN-0002	352070	11542	Breast cancer	Self-reported history of breast cancer, or cancer registration or death due to ICD10 C50, D48.6
CAN-0003	352070	15901	Skin cancer, including melanoma	Self-reported history of skin cancer, malignant melanoma, non-melanoma skin cancer, basal cell carcinoma, or squamous cell carcinoma, or cancer registration or death due to ICD10 C43-44, C79.2, D48.5
CAN-0004	352070	1966	Cancer of lip/mouth/pharynx/larynx/oral cavity	Self-reported history of cancer of lip/mouth/pharynx/oral cavity, salivary gland cancer, larynx/throat cancer, lip cancer, tongue cancer, gum cancer, parotid gland cancer, other salivary gland cancer, rodent ulcer, mouth cancer, tonsil cancer or oropharynx / oropharyngeal cancer, or cancer registration or death due to ICD10 C00-C14, C32, D00.0, D02.0, D38.0, D37.0
CAN-0005	352070	653	Oesophageal cancer	Self-reported history of oesophageal cancer, or cancer registration or death due to ICD10 C15, D00.1
CAN-0006	352070	619	Stomach cancer	Self-reported history of stomach cancer, or cancer registration or death due to ICD10 C16, D00.2, D37.1
CAN-0008	352070	14263	Intestinal tract cancer	Self-reported history of small intestine/small bowel cancer, large bowel cancer/colorectal cancer, anal cancer, colon cancer/sigmoid cancer, rectal cancer or appendix cancer, or cancer registration or death due to ICD10 C17-C21, C26.0, C78.4-C78.5, D37.2-D37.5, D01.0-D01.4, D12
CAN-0009	352070	51514	All cancer	Self-reported history of all types of cancers, or cancer registration or death due to ICD10 C*
CAN-0010	352070	2254	Liver cancer	Self-reported history of liver/hepatocellular cancer, or cancer registration or death due to ICD10 C78.7, D01.5, D37.6
CAN-0012	352070	570	Pancreas cancer	Self-reported history of pancreas cancer, or malignant insulinoma, or cancer registration or death due to ICD10 C25, D13.6
CAN-0013	352070	1098	Mesothelial, connective and soft tissue cancer	Self-reported history of peripheral nerve/autonomic nerve cancer, mesothelioma, sarcoma/fibrosarcoma, or kaposis sarcoma, or cancer registration or death due to ICD10 C21, C45-C49, D48.1-D48.4, D78.6

CAN-0015	352070	1591	Brain/central nervous cancer	Self-reported history of meningeal cancer / malignant meningioma, brain cancer / primary malignant brain tumour, or spinal cord or cranial nerve cancer, or cancer registration or death due to ICD10 C70-C72, C79.3, D33, D43
CAN-0016	352070	3362	Urinary cancer	Self-reported history of kidney/renal cell cancer, bladder cancer, or other cancer of urinary tract, or cancer registration or death due to ICD10 C64-C68, C79.0, D41.0-D41.4, D41.9, D09.0-D09.1, D30.3
CAN-0017	189755	1233	Ovary cancer	Self-reported history of ovarian cancer, or cancer registration or death due to ICD10 C56, D39.1
CAN-0018	352070	1603	Uterine/endometrical cancer	Self-reported history of uterine/endometrial cancer, or cancer registration or death due to ICD10 C54-C55, D39.0, D07.0
CAN-0019	189755	3251	Cervical cancer	Self-reported history of cervical cancer, or cin/pre-cancer cells cervix, or cancer registration or death due to ICD10 C53, D06
CAN-0023	189755	6088	Female genital cancer	Self-reported history of ovarian cancer, uterine/endometrial cancer, cervical cancer, or cin/pre-cancer cells cervix, vaginal cancer, vulval cancer, or female genital tract cancer, or cancer registration or death due to ICD10 C51-C57, D39.0-D39.2, D39.7, D06, D07.0-D07.3
CAN-0024	162291	5737	Prostate cancer	Self-reported history of prostate cancer, or cancer registration or death due to ICD10 C61, D07.5, D40.0
CAN-0025	162291	708	Testicular cancer	Self-reported history of testicular cancer, or cancer registration or death due to ICD10 C62, D40.1
CAN-0027	162291	6518	Male genital cancer	Self-reported history of prostate cancer, testicular cancer, penis cancer, or male genital tract cancer, or cancer registration or death due to ICD10 C60-C63, D07.4-D07.6, D40.0-D40.1, D40.7
CAN-0028	352070	2093	Lymphoma	Self-reported history of lymphoma, hodgkins lymphoma / hodgkins disease, or non-hodgkins lymphoma, or cancer registration or death due to ICD10 C81-C85, C88
CAN-0029	352070	1472	Leukaemia	Self-reported history of leukaemia, multiple myeloma, myelofibrosis or myelodysplasia, chronic lymphocytic, chronic myeloid, or acute myeloid leukaemia, or cancer registration or death due to ICD10 C90-C95

CAN-0030	352070	5405	Haematological malignancy	Self-reported history of lymphoma, hodgkins lymphoma / hodgkins disease, non-hodgkins lymphoma, leukaemia, multiple myeloma, myelofibrosis or myelodysplasia, chronic lymphocytic, chronic myeloid, acute myeloid leukaemia, or other haematological malignancy, or cancer registration or death due to ICD10 C81-C85, C88, C90-C96, C42.0-C42.1, C42.4, D18, D45-D47
CAN-0032	352070	680	Thyroid cancer	Self-reported history of thyroid cancer, or cancer registration or death due to ICD10 C73, D34, D44.0
CAN-0034	352070	1611	Endocrine gland cancer	Self-reported history of thyroid cancer, adrenal cancer, or parathyroid cancer, or cancer registration or death due to ICD10 C73-C75, D09.3, D34, D35, D44, D79.7
CAN-0036	352070	5655	Metastatic(secondary) cancer	Self-reported history of metastatic cancer (unknown primary), or cancer registration or death due to ICD10 C76, C78-C80
CAN-0038	352070	634	Heart/mediastinum/pleura cancer	Self-reported history of heart/mediastinum cancer, or cancer registration or death due to ICD10 C38, C45.0, C78.2

Supplementary Table 17: Diseases associated with genetically predicted LTL at Bonferroni-corrected significance (*p*-value (IVW) < $1.3X10^{-4}$).

Disease description				IVW-MR			Cochran's Q test		MR Egger			
Disease_super_category	Disease_ID	Disease_Name	N_case	Beta	SE	p-value	Cochran's Q	p-value	Beta	SE	p-value	intercept p-value
Neoplasms	ICD10-D25	Leiomyoma of uterus	5333	1.01	0.12	4.07E-18	33.38	0.01	1.73	0.34	3.11E-07	0.21
	ICD10-D17	Benign lipomatous neoplasm	3818	0.83	0.14	1.23E-09	23.89	0.12	1.81	0.40	6.27E-06	0.07
	ICD10-D22	Melanocytic naevi	2850	0.65	0.16	4.09E-05	18.75	0.34	1.75	0.47	2.02E-04	0.02
	CAN-0009	all cancer	51514	0.48	0.04	7.92E-33	13.76	0.68	0.82	0.12	3.63E-12	0.12
	CAN-0030	haematological malignancy	5405	0.88	0.12	2.35E-14	41.65	0.00	2.11	0.34	3.98E-10	0.04
	CAN-0029	leukaemia	1472	1.26	0.22	1.05E-08	20.31	0.26	2.47	0.65	1.34E-04	0.03
	CAN-1061	basal cell carcinoma	3399	0.68	0.14	2.56E-06	22.17	0.18	1.58	0.43	2.06E-04	0.08
	CAN-0024	prostate cancer	5737	0.79	0.11	5.77E-12	37.12	0.00	0.62	0.34	6.82E-02	0.81
	CAN-0027	male genital cancer	6518	0.61	0.11	1.08E-08	34.30	0.01	0.49	0.32	1.17E-01	0.85
	CAN-0034	endocrine gland cancer	1611	0.98	0.21	2.80E-06	5.11	1.00	1.36	0.62	2.82E-02	0.26
	CAN-0032	thyroid cancer	680	1.39	0.33	1.86E-05	17.88	0.40	2.55	0.98	9.50E-03	0.15
	CAN-0016	urinary cancer	3362	0.72	0.15	8.24E-07	32.43	0.01	1.81	0.42	1.81E-05	0.07
	CAN-0003	skin cancer, including melanoma	15901	0.60	0.07	9.75E-19	23.41	0.14	0.96	0.20	1.34E-06	0.34
	CAN-1059	malignant melanoma	2869	0.96	0.16	8.34E-10	24.95	0.10	1.28	0.46	5.44E-03	0.62
	ICD10-C44	Other malignant neoplasms of skin	4934	0.54	0.12	7.85E-06	30.47	0.02	0.55	0.35	1.10E-01	0.97

Extra columns are shown on the next page

Disease description			Median-MR			Penalised median-MR		
Disease_super_category	Disease_ID	Disease_Name	Beta	SE	P value	Beta	SE	p-value
Neoplasms	ICD10-D25	Leiomyoma of uterus	1.29	0.19	9.69E-12	1.30	0.18	1.68E-12
	ICD10-D17	Benign lipomatous neoplasm	0.93	0.23	3.83E-05	0.78	0.22	4.24E-04
	ICD10-D22	Melanocytic naevi	0.83	0.24	5.48E-04	0.84	0.24	4.88E-04
	CAN-0009	all cancer	0.45	0.06	5.65E-12	0.43	0.06	2.46E-11
	CAN-0030	haematological malignancy	0.95	0.21	4.96E-06	0.74	0.20	2.67E-04
	CAN-0029	leukaemia	1.41	0.30	2.27E-06	1.41	0.31	5.22E-06
	CAN-1061	basal cell carcinoma	0.72	0.22	1.10E-03	0.71	0.22	1.20E-03
	CAN-0024	prostate cancer	0.71	0.21	5.57E-04	0.72	0.20	3.47E-04
	CAN-0027	male genital cancer	0.58	0.18	1.39E-03	0.66	0.18	3.03E-04
	CAN-0034	endocrine gland cancer	1.14	0.28	5.32E-05	1.14	0.30	1.35E-04
	CAN-0032	thyroid cancer	1.58	0.50	1.54E-03	1.58	0.48	8.74E-04
	CAN-0016	urinary cancer	1.04	0.23	8.71E-06	1.02	0.22	4.05E-06
	CAN-0003	skin cancer, including melanoma	0.45	0.11	5.17E-05	0.42	0.12	2.27E-04
	CAN-1059	malignant melanoma	0.70	0.23	2.05E-03	0.67	0.23	3.06E-03
	ICD10-C44	Other malignant neoplasms of skin	0.34	0.18	6.83E-02	0.30	0.19	1.24E-01

Extra rows are shown on the next page

Disease description				IVW-MR			Cochran's Q test		MR Egger			
Disease_super_category	Disease_ID	Disease_Name	N_case	Beta	SE	p-value	Cochran's Q	p-value	Beta	SE	p-value	intercept p-value
Factors influencing health status and contact with health services	ICD10-Z46	Fitting and adjustment of other devices	2314	0.70	0.17	6.34E-05	18.39	0.36	0.49	0.51	3.28E-01	0.71
	ICD10-Z85	Personal history of malignant neoplasm	8296	0.42	0.09	5.88E-06	14.52	0.63	0.97	0.27	3.84E-04	0.09
Diseases of the genitourinary system	ICD10-N40	Hyperplasia of prostate	4727	0.81	0.12	5.90E-11	38.18	0.00	1.01	0.36	4.68E-03	0.78
	ICD10-N42	Other disorders of prostate	557	1.45	0.36	4.98E-05	29.24	0.03	2.08	1.04	4.61E-02	0.54
	ICD10-N84	Polyp of female genital tract	5905	0.56	0.11	4.89E-07	6.58	0.99	0.86	0.32	7.96E-03	0.22
Diseases of the digestive system	ICD10-K90	Intestinal malabsorption	987	-1.19	0.26	5.53E-06	23.33	0.14	-1.68	0.73	2.21E-02	0.49
Diseases of the circulatory system	ICD10-I10	Essential (primary) hypertension	29330	0.20	0.05	7.97E-05	7.34	0.98	0.53	0.15	3.67E-04	0.04

Median-MR			Penalised median-MR		
Beta	SE	P value	Beta	SE	p-value
0.77	0.26	2.46E-03	0.78	0.26	2.40E-03
0.45	0.14	1.04E-03	0.45	0.14	1.08E-03
0.86	0.21	5.11E-05	0.87	0.21	5.31E-05
1.50	0.50	2.79E-03	1.49	0.49	2.56E-03
0.56	0.15	1.59E-04	0.57	0.15	2.00E-04
-1.12	0.35	1.46E-03	-1.14	0.35	1.11E-03
0.16	0.08	3.48E-02	0.17	0.08	3.19E-02

Supplementary Table 18: Distribution of the mLRRY values in each EPIC-InterAct country, separately, and overall. Distributions before (upper) and after (bottom) data transformation are shown.

Country	N	Mean (SD)	Skewness	Kurtosis	Minimum	Maximum
Raw measurements of mLOY						
UK Biobank						
Overall	221,597	0 (0.05)	-1.22	7.31	-0.31	0.29
InterAct						
Overall	6,099	0.01 (0.08)	-23.11	935.34	-3.77	0.33
country						
Italy	492	0.02 (0.04)	-2.17	25.62	-0.35	0.20
Spain	882	0.03 (0.04)	-2.58	40.99	-0.49	0.28
UK	453	0.01 (0.10)	-9.46	130.47	-1.57	0.24
Netherlands	227	0.02 (0.05)	-5.06	51.32	-0.52	0.16
Germany	887	0.02 (0.05)	-4.10	46.46	-0.61	0.25
Sweden	1,093	0 (0.14)	-19.80	491.89	-3.77	0.33
Denmark	2,065	0.01 (0.05)	-10.40	172.36	-1.15	0.15
After winsorisation and rank-based inverse normal transformation						
UK Biobank						
Overall	221,597	0 (1.00)	0.00	3.00	-4.41	4.41
InterAct						
Overall	6,073	0 (1.00)	0.00	2.95	-3.42	3.42
Italy	491	0.13 (0.94)	0.07	2.71	-2.65	2.81
Spain	881	0.40 (0.94)	-0.15	3.34	-3.40	3.09
UK	450	0.08 (1.12)	-0.34	2.99	-3.21	3.23
Netherlands	226	0.35 (1.02)	-0.24	3.02	-2.85	3.42
Germany	884	0.16 (1.01)	-0.05	2.87	-3.00	2.93
Sweden	1,085	-0.30 (0.97)	0.25	3.19	-3.02	3.40
Denmark	2,056	-0.17 (0.93)	0.02	3.04	-3.42	3.11

Supplementary Table 19: Observational associations between mLOY (binary, mLRRY<0) and T2D risk in UK Biobank. Associations were analysed using logistic or Cox regression models for prevalent and incident T2D cases, respectively, with different adjustments, as shown in the table.

	Incident T2D (Cox regression models)						Prevalent T2D (logistic regression models)					
Additional adjustment	**HR**	**Beta**	**SE**	**P-value**	**total N**	**case N**	**OR**	**Beta**	**SE**	**P-value**	**n_total**	**n_case**
centre, array	1.09	0.09	0.03	7.58E-04	196,171	6,831	0.97	-0.03	0.02	8.95E-02	218,665	12,490
age, centre, array	1.08	0.08	0.03	2.26E-03	196,171	6,831	1.12	0.12	0.02	5.46E-10	218,665	12,490
age, smoking, centre, array	1.10	0.09	0.03	4.80E-04	195,992	6,822	1.13	0.12	0.02	1.34E-10	218,428	12,462
age, smoking, alchol consumption, education, BMI and wasit circumference, centre, array	1.07	0.06	0.03	1.59E-02	195,172	6,757	1.10	0.09	0.02	1.69E-06	217,239	12,357

Supplementary Table 20: Age or smoking stratification analyses in UK Biobank. Associations were analysed using logistic or Cox regression models for prevalent and incident T2D cases, respectively. Models were adjusted for centre and array in the age-band stratified analyses, and additionally for age in the smoking stratified analyses.

Incident T2D	Age band (years)	HR	Beta	SE	P-value	n_total	n_case
	<50	1.08	0.07	0.04	8.84E-02	47,313	755
	50-59	0.99	-0.01	0.02	8.05E-01	63,040	1,818
	60-69	1.08	0.07	0.02	2.93E-06	84,812	4,194
	70-73	1.12	0.11	0.13	3.91E-01	1,006	64
	$P_{interaction}$ = 0.4286						
	Never smoker	1.09	0.09	0.02	1.29E-04	97,300	2,455
	Ever smoker	1.05	0.05	0.02	1.73E-03	97,913	4,324
	$P_{interaction}$ = 0.1106						
Prevalent T2D	**Age band (years)**	**OR**	**Beta**	**SE**	**P-value**	**n_total**	**n_case**
	<50	1.11	0.11	0.03	0.001	51,623	1,124
	50-59	1.04	0.03	0.02	0.058	70,960	3,543
	60-69	1.06	0.06	0.01	5.14E-07	97,841	7,700
	70-73	1.32	0.28	0.09	0.002	1,167	123
	$P_{interaction}$ = 0.0208						
	Never smoker	1.10	0.09	0.02	9.43E-10	107,105	4,666
	Ever smoker	1.08	0.08	0.01	7.42E-11	110,401	7,697
	$P_{interaction}$ = 0.1286						

Supplementary Table 21: Meta-regression analyses to identify sources of heterogeneity for associations between mLRRY and T2D risk. Smoking status and age band were analysed separately, i.e. individuals were stratified by country and smoking status (ever vs. never) or by country and age band (<50, 50-65 and >65), resulting in 14 and 18 strata, respectively. In each stratified analysis, beta coefficients were combined across strata using random-effects meta-regression models. variances between strata (tau^2) were estimated by the residual (restricted) maximum likelihood (REML) algorithm with Knapp and Hartung modification to control type I error. In addition, permutation-based *p*-values were calculated, either with or without adjustment for multiple testing.

	REML method							
Study-level covariate	Beta	Se	T stat	KH-adjusted p	Tau^2_total	Tau^2_res	R^2(%)	I^2_res (%)
Country	0.04	0.01	4.30	1.00E-03	0.02	0.00	100.00	13.95
Age (50-65 *vs.* <50 ys)	-0.11	0.07	-1.54	0.15				
Age (>65 *vs.* <50 ys)	4.47E-03	0.08	0.06	0.96				
Country	0.04	0.01	5.09	3.58E-07	0.01	0.00	100.00	0.00
Smoking	-0.01	0.05	-0.11	0.91				

Moment-based method with permutation (5,000)				
T stat (observed)	p	p-joint	p-adjusted	N_studies
3.97	4.00E-03		0.01	18
-1.52	0.18	0.36	0.40	
0.11	0.92		1.00	
5.09	2.00E-04		2.00E-04	14
-0.11	0.91		0.99	

Supplementary Table 22: Missingness of mLRRY in EPIC-InterAct study. **A**. Proportion of mLRRY missingness in each country. **B**. Proportion of mLRRY missingness in T2D incident cases and controls. **C**. Age distribution among individuals with or without mLOY measurements. **D**. Factors associated with the missingness of mLRRY.

A.

country	mLRRY		Total (missing %)
	non-missing	*missing*	
Italy	492	752	1244 (60.45)
Spain	882	1,813	2695 (67.27)
UK	453	625	1078 (57.98)
Netherlands	227	220	447 (49.22)
Germany	887	979	1866 (52.47)
Sweden	1,093	1,541	2634 (58.5)
Denmark	2,065	209	2274 (9.19)
Total	6,099	6,139	12238 (50.16)

B.

T2D status	mLRRY		Total (missing %)
	non-missing	*missing*	
control	3,035	2,692	5,727 (47.01)
case	2,805	3,360	6,165 (54.50)

C.

mLRRY	Age		Total
	Mean	*SD*	
non-missing	54.49	8.11	6,095
missing	53.75	8.56	6,108
total	54.12	8.35	12,203

D. Missingness of mLRRY was coded as a binary variable: non-missing = 0, missing = 1. Logistic regression testing conditional effects of age, T2D status and country on the missingness of mLRRY.

	chi-sqaure	df	P-value
Age	4.67	1	0.0306
T2D	81.26	1	1.9789E-19
Country	1424.97	6	9.494E-305

Supplementary References

1. Singh, P. P., Demmitt, B. A., Nath, R. D. & Brunet, A. The Genetics of Aging: A Vertebrate Perspective. *Cell* **177**, 200–220 (2019).
2. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, (2013).
3. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
4. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
5. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, (2017).
6. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).
7. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).
8. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-.).* **291**, 1304 LP – 1351 (2001).
9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **304**, 1321 LP – 1325 (2004).
11. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
12. Altshuler, D., Donnelly, P. & Consortium, T. I. H. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
13. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
14. Consortium, the H. R. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279 (2016).
15. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
16. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–50 (2014).
17. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).
18. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
19. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
20. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
21. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
22. Wray, N. R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res. Hum. Genet.* **8**, 87–94

(2005).
23. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
24. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).
25. Chen, B., Cole, J. W. & Grond-Ginsbach, C. Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Front. Genet.* **8**, 167 (2017).
26. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
27. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
28. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
29. Moskvina, V. & Schmidt, K. M. On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.* **32**, 567–573 (2008).
30. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
31. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
32. Li, M.-X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).
33. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
34. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
35. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
36. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.* **96**, 434–42 (2004).
37. Wakefield, J. A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
38. DeWan, A. *et al.* HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration. *Science (80-.).* **314**, 989 LP – 992 (2006).
39. Consortium, T. W. T. C. C. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 (2007).
40. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, (2018).
41. Marigorta, U. M. & Navarro, A. High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLOS Genet.* **9**, e1003566 (2013).
42. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255 (2008).
43. Mackay, T. F. The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**, 303–339 (2001).
44. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–7

(2016).
45. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17**, 782 (2014).
46. Polychronakos, C. & Li, Q. Understanding type 1 diabetes through genetics: advances and prospects. *Nat. Rev. Genet.* **12**, 781–792 (2011).
47. Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLoS Genet.* **7**, e1002293 (2011).
48. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110 (2017).
49. Jiang, X. *et al.* Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat. Commun.* **9**, 260 (2018).
50. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet (London, England)* **376**, 180–8 (2010).
51. Wu, M. C. *et al.* Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am. J. Hum. Genet.* **86**, 929–942 (2010).
52. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
53. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *bioRxiv* (2019).
54. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
55. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
56. Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). The ENCODE Project Consortium. *PLoS Biol.* **9**, e1001046 (2011).
57. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
58. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
59. GTEx consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
60. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
61. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–52 (2016).
62. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **0**, 473–487 (2017).
63. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
64. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
65. Bothwell, L. E., Greene, J. A., Podolsky, S. H. & Jones, D. S. Assessing the Gold Standard — Lessons from the History of RCTs. *N. Engl. J. Med.* **374**, 2175–2181 (2016).

66. Altman, N. & Krzywinski, M. Association, correlation and causation. *Nat. Methods* **12**, 899–900 (2015).
67. Davies, N. M., Holmes, M. V & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
68. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
69. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Stat. Med.* **35**, 1880–1906 (2016).
70. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).
71. Pierce, B. L. & Burgess, S. Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).
72. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).
73. Swerdlow, D. I. *et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45**, 1600–1616 (2016).
74. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–14 (2016).
75. Maas, P. *et al.* Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol.* **2**, 1295 (2016).
76. Desikan, R. S. *et al.* Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLOS Med.* **14**, e1002258 (2017).
77. Seibert, T. M. *et al.* Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, j5757 (2018).
78. Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N. Engl. J. Med.* **375**, 2349–2358 (2016).
79. Ibanez, L., Farias, F. H. G., Dube, U., Mihindukulasuriya, K. A. & Harari, O. Polygenic Risk Scores in Neurodegenerative Diseases: a Review. *Curr. Genet. Med. Rep.* **7**, 22–29 (2019).
80. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
81. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
82. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013).
83. Richardson, T. G., Harrison, S., Hemani, G. & Davey Smith, G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* **8**, (2019).
84. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

85. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
86. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).
87. McCarthy, M. I. & Mahajan, A. The value of genetic risk scores in precision medicine for diabetes. *Expert Rev. Precis. Med. Drug Dev.* **3**, 279–281 (2018).
88. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
89. Meigs, J. B. *et al.* Genotype Score in Addition to Common Risk Factors for Prediction of Type 2 Diabetes. *N. Engl. J. Med.* **359**, 2208–2219 (2008).
90. Lango, H. *et al.* Assessing the Combined Impact of 18 Common Genetic Variants of Modest Effect Sizes on Type 2 Diabetes Risk. *Diabetes* **57**, 3129–3135 (2008).
91. van Hoek, M. *et al.* Predicting Type 2 Diabetes Based on Polymorphisms From Genome-Wide Association Studies: A Population-Based Study. *Diabetes* **57**, 3122–3128 (2008).
92. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
93. World Health Organization. World Report on Ageing and Health. *www.who.int* (2015).
94. Samani, N. J. & van der Harst, P. Biological ageing and cardiovascular disease. *Heart* **94**, 537–9 (2008).
95. Finkel, T., Serrano, M. & Blasco, M. A. The common biology of cancer and ageing. *Nature* **448**, 767–774 (2007).
96. Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNf676 as genes regulating telomere homeostasis in humans. *Hum. Mol. Genet.* **21**, 5385–5394 (2012).
97. McDaid, A. F. *et al.* Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.* **8**, 15842 (2017).
98. Graham Ruby, J. *et al.* Estimates of the heritability of human longevity are substantially inflated due to assortative mating. *Genetics* **210**, 1109–1124 (2018).
99. Partridge, L. & Gems, D. Mechanisms of ageing: public or private? *Nat. Rev. Genet.* **3**, 165–75 (2002).
100. Christensen, K., Johnson, T. E. & Vaupel, J. W. The quest for genetic determinants of human longevity: challenges and insights. *Nat. Rev. Genet.* **7**, 436–448 (2006).
101. Kenyon, C. J. The genetics of ageing. *Nature* **464**, 504–512 (2010).
102. Johnson, S. C., Rabinovitch, P. S. & Kaeberlein, M. mTOR is a key modulator of ageing and age-related disease. *Nature* **493**, 338–345 (2013).
103. Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
104. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A C. elegans mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
105. Kimura, K. D., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. daf-2, an Insulin Receptor-Like Gene That Regulates Longevity and Diapause in Caenorhabditis elegans. *Science (80-.).* **277**, 942–946 (1997).
106. Milman, S. *et al.* Low insulin-like growth factor-1 level predicts survival in humans with exceptional longevity. *Aging Cell* **13**, 769–771 (2014).
107. Suh, Y. *et al.* Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc. Natl. Acad. Sci.* **105**, 3438–3442 (2008).

108. Martins, R., Lithgow, G. J. & Link, W. Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell* **15**, 196–207 (2016).
109. Flachsbart, F. *et al.* Association of FOXO3A variation with human longevity confirmed in German centenarians. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2700–5 (2009).
110. Youngman, L. *et al.* Protein oxidation associated with aging is reduced by dietary restriction of protein or calories. *PNAS* **89**, 9112–9116 (2008).
111. Broer, L. *et al.* GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **70**, 110–118 (2015).
112. Flachsbart, F. *et al.* Identification and characterization of two functional variants in the human longevity gene FOXO3. *Nat. Commun.* **8**, 2063 (2017).
113. Sun, L. Y. *et al.* Longevity is impacted by growth hormone action during early postnatal period. *Elife* **6**, e24059 (2017).
114. Ben-Avraham, D. *et al.* The GH receptor exon 3 deletion is a marker of male-specific exceptional longevity associated with increased GH sensitivity and taller stature. *Sci. Adv.* **3**, e1602025 (2017).
115. van den Berg, N., Beekman, M., Smith, K. R., Janssens, A. & Slagboom, P. E. Historical demography and longevity genetics: Back to the future. *Ageing Res. Rev.* **38**, 28–39 (2017).
116. Giuliani, C., Garagnani, P. & Franceschi, C. Genetics of Human Longevity Within an Eco-Evolutionary Nature-Nurture Framework. *Circ. Res.* **123**, 745–772 (2018).
117. Albani, D. *et al.* Modulation of human longevity by SIRT3 single nucleotide polymorphisms in the prospective study “Treviso Longeva (TRELONG)”. *Age (Omaha).* **36**, 469–478 (2014).
118. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* **8**, 1–40 (2019).
119. Joshi, P. K. *et al.* Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nat. Commun.* **8**, 1–13 (2017).
120. Joshi, P. K. *et al.* Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nat. Commun.* **7**, 11174 (2016).
121. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing. *Nature* **561**, 45–56 (2018).
122. Scaffidi, P. & Misteli, T. Lamin A-dependent nuclear defects in human aging. *Science* **312**, 1059–63 (2006).
123. De Sandre-Giovannoli, A. *et al.* Lamin a truncation in Hutchinson-Gilford progeria. *Science* **300**, 2055 (2003).
124. Zhang, W. *et al.* Aging stem cells. A Werner syndrome stem cell model unveils heterochromatin alterations as a driver of human aging. *Science* **348**, 1160–3 (2015).
125. Kudlow, B. A., Kennedy, B. K. & Jr, R. J. M. Werner and Hutchinson – Gilford progeria syndromes : mechanistic basis of human progeroid diseases. *Nat. Rev. Mol. Cell Biol.* **8**, 394–404 (2007).
126. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23**, 4420–4432 (2014).
127. Newman, A. B. *et al.* A Meta-analysis of four genome-wide association studies of survival to age 90 years or older: The cohorts for heart and aging research in genomic epidemiology consortium. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **65 A**, 478–487 (2010).

128. Broer, L. *et al.* Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur. J. Hum. Genet.* **21**, 1163–1168 (2013).
129. Walter, S. *et al.* A genome-wide association study of aging. *Neurobiol. Aging* **32**, (2011).
130. Pilling, L. C., Atkins, J. L., Bowman, K. & Jones, S. E. longevity is influenced by many genetic variants : evidence from 75 , 000 UK Biobank participants. *Aging (Albany. NY).* **8**, 547–560 (2016).
131. Fortney, K. *et al.* Genome-Wide Scan Informed by Age-Related Disease Identifies Loci for Exceptional Human Longevity. *PLoS Genet.* **11**, 1–23 (2015).
132. Deelen, J. *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, (2019).
133. Medici, M. *et al.* Identification of Novel Genetic Loci Associated with Thyroid Peroxidase Antibodies and Clinical Thyroid Disease. *PLoS Genet.* **10**, e1004123 (2014).
134. Jazwinski, S. M. & Kim, S. Examination of the dimensions of biological age. *Frontiers in Genetics* **10**, (2019).
135. Thompson, D. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood is associated with genomic instability in other tissues and susceptibility to non-haematological cancers. *bioRxiv* 514026 (2019). doi:10.1101/514026
136. Sebastiani, P. *et al.* Biomarker signatures of aging. *Aging Cell* **16**, 329–338 (2017).
137. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **8**, 14617 (2017).
138. Lu, A. T. *et al.* GWAS of epigenetic aging rates in blood reveals a critical role for TERT. *Nat. Commun.* **9**, (2018).
139. Suzuki, M. M. & Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
140. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147 (2018).
141. de Lange, T. How Telomeres Solve the End-Protection Problem. *Science (80-.).* **326**, 948–952 (2009).
142. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat. Med.* **12**, 1133–1138 (2006).
143. Blasco, M. A. The epigenetic regulation of mammalian telomeres. *Nat Rev Genet* **8**, 299–309 (2007).
144. Fagagna, F. d'Adda di *et al.* A DNA damage checkpoint response in telomere-initiated senescence. *Nature* **426**, 194–198 (2003).
145. Takai, H., Smogorzewska, A. & de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr. Biol.* **13**, 1549–56 (2003).
146. Allsopp, R. C. *et al.* Telomere length predicts replicative capacity of human fibroblasts. *Proc. Natl. Acad. Sci.* **89**, 10114–10118 (1992).
147. O'Sullivan, R. J. & Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat. Rev. Mol. Cell Biol.* **11**, 171 (2010).
148. Wang, C. & Meier, U. T. Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J.* **23**, 1857–1867 (2004).
149. Bischoff, C. *et al.* The Heritability of Telomere Length Among the Elderly and Oldest-Old. *Twin Res Hum Genet* **8**, 433–439 (2005).

150. Vasa-nicotera, M. *et al.* Mapping of a Major Locus that Determines Telomere Length in Humans. *Am. J. Hum. Genet.* **76**, 147–151 (2005).
151. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.* **45**, 422 (2013).
152. Codd, V. *et al.* Common variants near TERC are associated with mean telomere length. *Nat. Genet.* **42**, 197–199 (2010).
153. Mangino, M. *et al.* DCAF4, a novel gene associated with leucocyte telomere length. *J. Med. Genet.* **52**, 157–162 (2015).
154. Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: Identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.* **22**, 5056–5064 (2013).
155. Ding, H. *et al.* Regulation of Murine Telomere Length by Rtel : An Essential Gene Encoding a Helicase-like Protein. *Cell* **117**, 873–886 (2004).
156. Pooley, K. A. *et al.* Telomere Length in Prospective and Retrospective Cancer Case-Control Studies. **78**, 3170–3177 (2010).
157. Gu, J. *et al.* A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte telomere length and as a marker of susceptibility for bladder cancer. *Cancer Prev. Res.* **4**, 514–521 (2011).
158. Haycock, P. C. *et al.* Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study. *J. Am. Med. Assoc. Oncol.* **3**, 636–651 (2017).
159. Shay, J. W. & Wright, W. E. Telomeres and telomerase: three decades of progress. *Nat. Rev. Genet.* **20**, 299–309 (2019).
160. Ryder, H. *et al.* Obesity , cigarette smoking , and telomere length in women. *Lancet* **366**, 662–664 (2005).
161. Müezzinler, A. *et al.* Body mass index and leukocyte telomere length dynamics among older adults : Results from the ESTHER cohort. *EXG* **74**, 1–8 (2016).
162. Wulaningsih, W., Kuh, D., Wong, A. & Hardy, R. Adiposity, Telomere Length, and Telomere Attrition in Midlife: the 1946 British Birth Cohort. *J Gerontol A Biol Sci Med Sci* **00**, 1–7 (2017).
163. Müezzinler, A. *et al.* Smoking habits and leukocyte telomere length dynamics among older adults : Results from the ESTHER cohort. *Exp Gerontol* **70**, 18–25 (2015).
164. Weischer, M., Bojesen, S. E. & Nordestgaard, B. G. Telomere Shortening Unrelated to Smoking , Body Weight , Physical Activity , and Alcohol Intake : 4 , 576 General Population Individuals with Repeat Measurements 10 Years Apart. *PLoS Genet* **10**, 1–11 (2014).
165. Angela R. Starkweather, PhD, ACNP-BC, CNRN, Areej A. Alhaeeri, BS, Alison Montpetit, PhD, RN, Jenni Brumelle, PhD, Kristin Filler, RN, BS, Marty Montpetit, PhD, Lathika Mohanraj, PhD, Debra E. Lyon, PhD, RN, FNP-BC, FNAP, FAAN, and C. K. J.-C. An Integrative Review of Factors Associated with Telomere Length and Implications for Biobehavioral Research. *Nurs Res* **100**, 130–134 (2014).
166. Lynn F. Cherkas *et al.* The Association Between Physical Activity in Leisure Time and Leukocyte Telomere Length. *J. Am. Med. Assoc.* **168**, 154–158 (2008).
167. Mundstock, E. *et al.* Effect of Obesity on Telomere Length : Systematic Review and. *Obesity* **23**, 2165–2174 (2015).
168. Adler, N. *et al.* NIH Public Access. *Brain Behav Immun* **27**, 15–21 (2014).
169. Kajantie, E. *et al.* No association between body size at birth and leucocyte telomere

length in adult life — evidence from three cohort studies. *Int. J Epidemiol* **41**, 1400–1408 (2012).
170. Theall, K. P., Shirtcliff, E. A., Dismukes, A. R., Wallace, M. & Drury, S. S. Association Between Neighborhood Violence and Biological Stress in Children. *J. Am. Med. Assoc. Pediatr.* **171**, 53–60 (2017).
171. Njajou, O. T. *et al.* Telomere length is paternally inherited and is associated with parental lifespan. **104**, 12135–12139 (2007).
172. Burtner, C. R. & Kennedy, B. K. Progeria syndromes and ageing: what is the connection? *Nat. Rev. Mol. Cell Biol.* **11**, 567–578 (2010).
173. Wong, J. M. Y. & Collins, K. Telomere maintenance and disease. *Lancet* **362**, 983–988 (2003).
174. Armanios, M. & Blackburn, E. H. The telomere syndromes. *Nat. Rev. Genet.* **13**, 693–704 (2012).
175. Howlett, N. G. Biallelic Inactivation of BRCA2 in Fanconi Anemia. *Science (80-.).* **297**, 606–609 (2002).
176. Blackburn, E. H., Epel, E. S. & Lin, J. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science (80-.).* **350**, 1193–1198 (2015).
177. Said, M. A., Eppinga, R. N., Hagemeijer, Y., Verweij, N. & van der Harst, P. Telomere Length and Risk of Cardiovascular Disease and Cancer. *J. Am. Coll. Cardiol.* **70**, 506–507 (2017).
178. Fyhrquist, F., Saijonmaa, O. & Strandberg, T. The roles of senescence and telomere shortening in cardiovascular disease. *Nat .Rev. Cardiol.* **10**, 274–283 (2013).
179. Haycock, P. C. *et al.* Leucocyte telomere length and risk of cardiovascular disease: systematic review and meta-analysis. *BMJ* **349**, (2014).
180. Zhan, Y. *et al.* Telomere Length Shortening and Alzheimer Disease—A Mendelian Randomization Study. *J. Am. Med. Assoc.* **72**, 1202–1203 (2015).
181. Honig, L. S., Kang, M. S., Schupf, N., Lee, J. H. & Mayeux, R. Association of Shorter Leukocyte Telomere Repeat Length With Dementia and Mortality. *J. Am. Med. Assoc. Neurol.* **69**, 1332 (2012).
182. D'Mello, M. J. J. *et al.* Association between shortened leukocyte telomere length and cardiometabolic outcomes: systematic review and meta-analysis. *Circ. Cardiovasc. Genet.* **8**, 82–90 (2015).
183. Forero, D. A. *et al.* Meta-analysis of Telomere Length in Alzheimer's Disease. *J. Gerontol. A. Biol. Sci. Med. Sci.* **71**, 1069–73 (2016).
184. Willeit, P., Willeit, J., Kloss-Brandstatter, Kronenberg, F. & Kiechl, S. Fifteen-Year Follow-up of Association Between Telomere Length and Incident Cancer and Cancer Mortality. *J. Am. Med. Assoc.* **306**, 42–44 (2011).
185. Willeit, P. *et al.* Telomere length and risk of incident cancer and cancer mortality. *J. Am. Med. Assoc.* **304**, 69–75 (2010).
186. Barthel, F. P. *et al.* Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* **49**, 349–357 (2017).
187. Graham, M. K. & Meeker, A. Telomeres and telomerase in prostate cancer development and therapy. *Nat. Rev. Urol.* **14**, 607–619 (2017).
188. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Publ. Gr.* **12**, 363–375 (2011).
189. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and

cancer. *Nat. Genet.* **44**, 651–658 (2012).
190. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
191. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
192. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
193. Forsberg, L. a *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
194. Loh, P. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
195. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
196. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–8 (2016).
197. Acuna-Hidalgo, R. *et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).
198. Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E. & Nickerson, D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40**, 1199–1203 (2008).
199. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
200. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease-clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
201. Vattathil, S. & Scheet, P. Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue. *Am. J. Hum. Genet.* **98**, 571–578 (2016).
202. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 1–7 (2016).
203. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
204. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
205. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
206. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
207. Dumanski, J. P. *et al.* Mosaic Loss of Chromosome y in Blood Is Associated with Alzheimer Disease. *Am. J. Hum. Genet.* **98**, 1208–1219 (2016).
208. Zhang, C. *et al.* Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Hum. Mol. Genet.* **24**, 5356–5366 (2015).
209. Iles, M. M. *et al.* The Effect on Melanoma Risk of Genes Previously Associated With Telomere Length. *JNCI J. Natl. Cancer Inst.* **106**, (2014).
210. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci. U.S.A* **107**, 9293–8 (2010).

211. Delgado, D. A. *et al.* Genome-wide association study of telomere length among South Asians identifies a second RTEL1 association signal. *J. Med. Genet.* **55**, 64–71 (2018).
212. Langenberg, C. *et al.* Design and cohort description of the InterAct Project: An examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* **54**, 2272–2282 (2011).
213. Langenberg, C. *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLoS Med.* **11**, (2014).
214. Danesh, J. *et al.* EPIC-Heart: The cardiovascular component of a prospective study of nutritional, lifestyle and biological factors in 520,000 middle-aged participants from 10 European countries. *Eur. J. Epidemiol.* **22**, 129–141 (2007).
215. Kristiansson, K. *et al.* Genome-Wide Screen for Metabolic Syndrome Susceptibility Loci Reveals Strong Lipid Gene Contribution But No Evidence for Common Genetic Basis for Clustering of Metabolic Syndrome Traits. *Circ. Cardiovasc. Genet.* **5**, 242–249 (2012).
216. Penninx, B. W. J. H. *et al.* The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121–140 (2008).
217. Ikram, M. A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* **32**, 807–850 (2017).
218. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
219. Storey, J. D. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **64**, 479–498 (2002).
220. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
221. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 1–22 (2013).
222. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
223. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
224. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
225. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
226. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
227. Wang, J., Dayem Ullah, A. Z. & Chelala, C. IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome. *Nucleic Acids Res.* **46**, e47–e47 (2018).
228. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
229. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
230. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
231. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation

across the human life course. *Genome Biol.* **17**, 61 (2016).
232. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
233. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, (2015).
234. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
235. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109-14 (2012).
236. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691-7 (2011).
237. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
238. Blake, J. A. *et al.* Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research* **45**, D723-9 (2017).
239. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
240. Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat. Commun.* **10**, (2019).
241. Krenciute, G. *et al.* Nuclear BAG6-UBL4A-GET4 Complex Mediates DNA Damage Signaling and Cell Death. *J. Biol. Chem.* **288**, 20547–20557 (2013).
242. Kim, Y. D. *et al.* Metformin Inhibits Hepatic Gluconeogenesis Through AMP-Activated Protein Kinase-Dependent Regulation of the Orphan Nuclear Receptor SHP. *Diabetes* **57**, 306–314 (2008).
243. Irwin, C. R., Hitt, M. M. & Evans, D. H. Targeting Nucleotide Biosynthesis: A Strategy for Improving the Oncolytic Potential of DNA Viruses. *Front. Oncol.* **7**, 229 (2017).
244. Reichard, P. Interactions between deoxyribonucleotide and DNA synthesis. *Annu. Rev. Biochem.* **57**, 349–374 (1988).
245. Pedroza-García, J. A. *et al.* Role of pyrimidine salvage pathway in the maintenance of organellar and nuclear genome integrity. *Plant J.* **97**, 430–446 (2019).
246. Echols, H. & Goodman, M. F. Fidelity mechanisms in DNA replication. *Annu. Rev. Biochem.* **60**, 477–511 (1991).
247. Bebenek, K., Roberts, J. D. & Kunkel, T. A. The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J. Biol. Chem.* **267**, 3589–3596 (1992).
248. Feng, I. J. & Radivoyevitch, T. SNP-SNP Interactions between dNTP Supply Enzymes and Mismatch DNA Repair in Breast Cancer. in *2009 Ohio Collaborative Conference on Bioinformatics* 123–128 (IEEE, 2009). doi:10.1109/OCCBIO.2009.25
249. Austin, W. R. *et al.* Nucleoside salvage pathway kinases regulate hematopoiesis by linking nucleotide metabolism with replication stress. *J. Exp. Med.* **209**, 2215 LP – 2228 (2012).
250. Franzolin, E. *et al.* The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14272–14277 (2013).
251. Jobert, L. *et al.* The Human Base Excision Repair Enzyme SMUG1 Directly Interacts with DKC1 and Contributes to RNA Quality Control. *Mol. Cell* **49**, 339–345 (2013).

252. de Lange, T. Shelterin-Mediated Telomere Protection. *Annu. Rev. Genet.* **52**, 223–247 (2018).
253. Deng, Z. *et al.* Inherited mutations in the helicase RTEL1 cause telomere dysfunction and Hoyeraal-Hreidarsson syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3408-16 (2013).
254. Giraud-Panis, M.-J., Teixeira, M. T., Géli, V. & Gilson, E. CST Meets Shelterin to Keep Telomeres in Check. *Mol. Cell* **39**, 665–676 (2010).
255. Kim, M. K. *et al.* Regulation of telomeric repeat binding factor 1 binding to telomeres by casein kinase 2-mediated phosphorylation. *J. Biol. Chem.* **283**, 14144–14152 (2008).
256. Lee, S. S., Bohrson, C., Pike, A. M., Wheelan, S. J. & Greider, C. W. ATM Kinase Is Required for Telomere Elongation in Mouse and Human Cells. *Cell Rep.* **13**, 1623–1632 (2015).
257. Tong, A. S. *et al.* ATM and ATR Signaling Regulate the Recruitment of Human Telomerase to Telomeres. *Cell Rep.* **13**, 1633–1646 (2015).
258. Beneke, S. *et al.* Rapid regulation of telomere length is mediated by poly(ADP-ribose) polymerase-1. *Nucleic Acids Res.* **36**, 6309–6317 (2008).
259. Gomez, M. *et al.* PARP1 Is a TRF2-associated poly(ADP-ribose)polymerase and protects eroded telomeres. *Mol. Biol. Cell* **17**, 1686–96 (2006).
260. Denchi, E. L. & de Lange, T. Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature* **448**, 1068–1071 (2007).
261. Karlseder, J., Broccoli, D., Dai, Y., Hardy, S. & de Lange, T. p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. *Science* **283**, 1321–1325 (1999).
262. van Steensel, B., Smogorzewska, A. & de Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
263. Arnoult, N. & Karlseder, J. Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol* **22**, 859–866 (2015).
264. Collins, K. & Mitchell, J. R. Telomerase in the human organism. *Oncogene* **21**, 564–579 (2002).
265. Blackburn, E. H. & Collins, K. Telomerase: An RNP Enzyme Synthesizes DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003558–a003558 (2011).
266. Stanley, S. E. *et al.* Loss-of-function mutations in the RNA biogenesis factor NAF1 predispose to pulmonary fibrosis-emphysema. *Sci. Transl. Med.* **8**, 351ra107 (2016).
267. Egan, E. D. & Collins, K. Biogenesis of telomerase ribonucleoproteins. *RNA* **18**, 1747–1759 (2012).
268. Nguyen, D. *et al.* A Polyadenylation-Dependent 3' End Maturation Pathway Is Required for the Synthesis of the Human Telomerase RNA. *Cell Rep.* **13**, 2244–57 (2015).
269. Moon, D. H. *et al.* Poly(A)-specific ribonuclease (PARN) mediates 3'-end maturation of the telomerase RNA component. *Nat. Genet.* **47**, 1482–1488 (2015).
270. Boyraz, B. *et al.* Posttranscriptional manipulation of TERC reverses molecular hallmarks of telomere disease. *J. Clin. Invest.* **126**, 3377–3382 (2016).
271. Deng, T. *et al.* TOE1 acts as a 3' exonuclease for telomerase RNA and regulates telomere maintenance. *Nucleic Acids Res.* **47**, 391–405 (2019).
272. Schilders, G., Raijmakers, R., Raats, J. M. H. & Pruijn, G. J. M. MPP6 is an exosome-associated RNA-binding protein involved in 5.8S rRNA maturation. *Nucleic Acids Res.* **33**, 6795–6804 (2005).

273. Arnér, E. S. & Eriksson, S. Mammalian deoxyribonucleoside kinases. *Pharmacol. Ther.* **67**, 155–86 (1995).
274. Mutahir, Z. *et al.* Thymidine kinase 1 regulatory fine-tuning through tetramer formation. *FEBS J.* **280**, 1531–1541 (2013).
275. Sabini, E., Hazra, S., Ort, S., Konrad, M. & Lavie, A. Structural basis for substrate promiscuity of dCK. *J. Mol. Biol.* **378**, 607–21 (2008).
276. Irwin, C. R., Hitt, M. M. & Evans, D. H. Targeting Nucleotide Biosynthesis: A Strategy for Improving the Oncolytic Potential of DNA Viruses. *Front. Oncol.* **7**, 229 (2017).
277. Carreras, C. W. & Santi, D. V. The Catalytic Mechanism and Structure of Thymidylate Synthase. *Annu. Rev. Biochem.* **64**, 721–762 (1995).
278. Anderson, D. D., Quintero, C. M. & Stover, P. J. Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. *Proc. Natl. Acad. Sci.* **108**, 15163 LP – 15168 (2011).
279. Bester, A. C. *et al.* Nucleotide Deficiency Promotes Genomic Instability in Early Stages of Cancer Development. *Cell* **145**, 435–446 (2011).
280. Chabes, A. *et al.* Survival of DNA Damage in Yeast Directly Depends on Increased dNTP Levels Allowed by Relaxed Feedback Inhibition of Ribonucleotide Reductase. *Cell* **112**, 391–401 (2003).
281. Davidson, M. B. *et al.* Endogenous DNA replication stress results in expansion of dNTP pools and a mutator phenotype. *EMBO J.* **31**, 895 LP – 907 (2012).
282. Blasco, M. A. Telomeres and human disease: ageing, cancer and beyond. *Nat. Rev. Genet.* **6**, 611–622 (2005).
283. Holohan, B., Wright, W. E. & Shay, J. W. Telomeropathies: An emerging spectrum disorder. *J. Cell Biol.* **205**, 289–299 (2014).
284. Sarek, G., Marzec, P., Margalef, P. & Boulton, S. J. Molecular basis of telomere dysfunction in human genetic diseases. *Nat. Struct. Mol. Biol* **22**, 867–874 (2015).
285. Brouilette, S. W. *et al.* Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study. *Lancet* **369**, 107–114 (2007).
286. Benetos, A. *et al.* Short Telomeres Are Associated With Increased Carotid Atherosclerosis in Hypertensive Subjects. *Hypertension* **43**, 182–185 (2004).
287. Brouilette, S., Singh, R. K., Thompson, J. R., Goodall, A. H. & Samani, N. J. White Cell Telomere Length and Risk of Premature Myocardial Infarction. *Arterioscler. Thromb. Vasc. Biol.* **23**, 842–846 (2003).
288. Fitzpatrick, A. L. *et al.* Leukocyte Telomere Length and Cardiovascular Disease in the Cardiovascular Health Study. *Am. J. Epidemiol.* **165**, 14–21 (2006).
289. Wentzensen, I. M., Mirabello, L., Pfeiffer, R. M. & Savage, S. A. The Association of Telomere Length and Cancer: a Meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **20**, 1238–1250 (2011).
290. Zhu, X. *et al.* The association between telomere length and cancer risk in population studies. *Sci. Rep.* **6**, 22243 (2016).
291. Zhan, Y. *et al.* Exploring the Causal Pathway From Telomere Length to Coronary Heart DiseaseNovelty and Significance. *Circ. Res.* **121**, 214–219 (2017).
292. PRENTICE, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11 (1986).
293. Morris, A., Voight, B., Teslovich, T., Ferreira, T. & Segre, A. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2

diabetes. **44**, (2012).
294. White, J. *et al.* Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. *JAMA Cardiol.* **366**, 1108–1118 (2016).
295. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
296. Prokopenko, I. *et al.* A Central Role for GRB10 in Regulation of Islet Function in Man. *PLoS Genet.* **10**, 1–13 (2014).
297. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–83 (2013).
298. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* **00**, 1–9 (2018).
299. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2018).
300. Zheng, J. *et al.* LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
301. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
302. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
303. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
304. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
305. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *Int. J. Epidemiol.* **32**, 1–22 (2003).
306. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
307. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
308. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
309. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. (2018).
310. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
311. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–

2376 (2014).
312. Staley, J. R. *et al.* PhenoScanner: A database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
313. Sanchez-Espiridion, B. *et al.* Telomere Length in Peripheral Blood Leukocytes and Lung Cancer Risk: A Large Case–Control Study in Caucasians. *Cancer Res.* **74**, 2476 LP – 2486 (2014).
314. Stone, R. C. *et al.* Telomere Length and the Cancer–Atherosclerosis Trade-Off. *PLOS Genet.* **12**, e1006144 (2016).
315. Savage, S. A., Gadalla, S. M. & Chanock, S. J. The Long and Short of Telomeres and Cancer Association Studies. *JNCI J. Natl. Cancer Inst.* **105**, 448–449 (2013).
316. McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406 (2008).
317. Speedy, H. E. *et al.* Germ line mutations in shelterin complex genes are associated with familial chronic lymphocytic leukemia. *Blood* **128**, 2319–2326 (2016).
318. Rode, L., Nordestgaard, B. G. & Bojesen, S. E. Long telomeres and cancer risk among 95 568 individuals from the general population. *Int. J. Epidemiol.* **45**, 1634–1643 (2016).
319. Landi, M. T. *et al.* A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679–691 (2009).
320. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **18**, 175–186 (2017).
321. Shi, J. *et al.* Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nat. Genet.* **46**, 482–486 (2014).
322. Weller, M. *et al.* Glioma. *Nat. Rev. Dis. Prim.* **1**, 15017 (2015).
323. Walsh, K. M. *et al.* Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk. *Oncotarget* **6**, 42468–77 (2015).
324. Holohan, B. *et al.* Decreasing initial telomere length in humans intergenerationally understates age-associated telomere shortening. *Aging Cell* **14**, 669–677 (2015).
325. Chen, W. *et al.* Longitudinal versus cross-sectional evaluations of leukocyte telomere length dynamics: age-dependent telomere shortening is the rule. *Journals Gerontol. Ser. A Biomed. Sci. Med. Sci.* **66**, 312–319 (2011).
326. DF, S., JA, B., SC, M. & al, et. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA* **283**, 2008–2012 (2000).
327. Cawthon, R. M. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.* **37**, e21–e21 (2009).
328. De Lucia Rolfe, E. *et al.* Association between birth weight and visceral fat in adults. *Am. J. Clin. Nutr.* **92**, 347–352 (2010).
329. Lindsay, T. *et al.* Descriptive epidemiology of physical activity energy expenditure in UK adults. The Fenland Study. *medRxiv* 19003442 (2019). doi:10.1101/19003442
330. Godino, J. G. *et al.* Effect of communicating genetic and phenotypic risk for type 2 diabetes in combination with lifestyle advice on objectively measured physical activity: protocol of a randomised controlled trial. *BMC Public Health* **12**, 444 (2012).
331. Cawthon, R. M. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.* **37**, 1–7 (2009).
332. Huzen, J. *et al.* Telomere length loss due to smoking and metabolic traits. *J. Intern. Med.* **275**, 155–163 (2014).

333. Dalgård, C. *et al.* Leukocyte telomere length dynamics in women and men: menopause vs age effects. *Int. J. Epidemiol.* **44**, 1688–1695 (2015).
334. Nordfjäll, K. *et al.* The individual blood cell telomere attrition rate is telomere length dependent. *PLoS Genet.* **5**, e1000375 (2009).
335. Verhulst, S., Aviv, A., Benetos, A., Berenson, G. S. & Kark, J. D. Do leukocyte telomere length dynamics depend on baseline telomere length? An analysis that corrects for 'regression to the mean'. *Eur. J. Epidemiol.* **28**, 859–866 (2013).
336. Farzaneh-Far, R. *et al.* Telomere length trajectory and its determinants in persons with coronary artery disease: longitudinal findings from the heart and soul study. *PLoS One* **5**, e8612 (2010).
337. Bendix, L. *et al.* Longitudinal changes in leukocyte telomere length and mortality in humans. *Journals Gerontol. Ser. A Biomed. Sci. Med. Sci.* **69**, 231–239 (2013).
338. Kark, J. D., Goldberger, N., Kimura, M., Sinnreich, R. & Aviv, A. Energy intake and leukocyte telomere length in young adults. *Am. J. Clin. Nutr.* **95**, 479–487 (2012).
339. Farzaneh-Far, R. *et al.* Association of marine omega-3 fatty acid levels with telomeric aging in patients with coronary heart disease. *JAMA* **303**, 250–257 (2010).
340. García-Calzón, S. *et al.* Dietary inflammatory index and telomere length in subjects with a high cardiovascular disease risk from the PREDIMED-NAVARRA study: cross-sectional and longitudinal analyses over 5 y. *Am. J. Clin. Nutr.* **102**, 897–904 (2015).
341. Eriksson, J. G. *et al.* Higher serum phenylalanine concentration is associated with more rapid telomere shortening in men. *Am. J. Clin. Nutr.* **105**, 144–150 (2016).
342. Soares-Miranda, L. *et al.* Physical Activity, Physical Fitness and Leukocyte Telomere Length: the Cardiovascular Health Study. *Med. Sci. Sports Exerc.* **47**, 2525 (2015).
343. Van Ockenburg, S. L., de Jonge, P., Van der Harst, P., Ormel, J. & Rosmalen, J. G. M. Does neuroticism make you old? Prospective associations between neuroticism and leukocyte telomere length. *Psychol. Med.* **44**, 723–729 (2014).
344. Van Ockenburg, S. L. *et al.* Stressful life events and leukocyte telomere attrition in adulthood: a prospective population-based cohort study. *Psychol. Med.* **45**, 2975–2984 (2015).
345. Dowd, J. B. *et al.* Persistent herpesvirus infections and telomere attrition over 3 years in the Whitehall II cohort. *J. Infect. Dis.* **216**, 565–572 (2017).
346. Ferreira, M. S. V. *et al.* Evidence for a pre-existing telomere deficit in non-clonal hematopoietic stem cells in patients with acute myeloid leukemia. *Ann. Hematol.* **96**, 1457–1461 (2017).
347. Townsley, D. M. *et al.* Danazol treatment for telomere diseases. *N. Engl. J. Med.* **374**, 1922–1931 (2016).
348. Ping, F. *et al.* Deoxyribonucleic acid telomere length shortening can predict the incidence of non-alcoholic fatty liver disease in patients with type 2 diabetes mellitus. *J. Diabetes Investig.* **8**, 174–180 (2017).
349. Masi, S. *et al.* Rate of telomere shortening and cardiovascular damage: a longitudinal study in the 1946 British Birth Cohort. *Eur. Heart J.* **35**, 3296–3303 (2014).
350. Epel, E. S. *et al.* The rate of leukocyte telomere shortening predicts mortality from cardiovascular disease in elderly men. *Aging (Albany NY)* **1**, 81 (2009).
351. Wang, L., Xiao, H., Zhang, X., Wang, C. & Huang, H. The role of telomeres and telomerase in hematologic malignancies and hematopoietic stem cell transplantation. *J. Hematol. Oncol.* **7**, 61 (2014).
352. Barnett, A. G., van der Pols, J. C. & Dobson, A. J. Regression to the mean: what it is

and how to deal with it. *Int. J. Epidemiol.* **34**, 215–220 (2004).
353. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).
354. Lleo, A. *et al.* Y chromosome loss in male patients with primary biliary cirrhosis. *J. Autoimmun.* **41**, 87–91 (2013).
355. Persani, L. *et al.* Increased loss of the Y chromosome in peripheral blood cells in male patients with autoimmune thyroiditis. *J. Autoimmun.* **38**, J193–J196 (2012).
356. Haitjema, S. *et al.* Loss of Y Chromosome in Blood Is Associated With Major Cardiovascular Events During Follow-Up in Men After Carotid Endarterectomy. *Circ. Cardiovasc. Genet.* **10**, (2017).
357. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
358. Zhou, W. *et al.* Reply to 'Mosaic loss of chromosome Y in leukocytes matters'. *Nat. Genet.* **51**, 7–9 (2019).
359. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).
360. Bonnefond, A. *et al.* Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
361. Zimmet, P., Alberti, K. G. M. M. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**, 782–787 (2001).
362. Barbieri, M., Bonafè, M., Franceschi, C. & Paolisso, G. Insulin/IGF-I-signaling pathway: an evolutionarily conserved mechanism of longevity from yeast to humans. *Am. J. Physiol. Metab.* **285**, E1064–E1071 (2003).
363. Tatar, M., Bartke, A. & Antebi, A. The Endocrine Regulation of Aging by Insulin-like Signals. *Science (80-.).* **299**, 1346–1351 (2003).
364. Abbasi, A. *et al.* Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* **345**, e5900 (2012).
365. Kengne, A. P. *et al.* Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): A validation of existing models. *Lancet Diabetes Endocrinol.* **2**, 19–29 (2014).
366. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
367. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
368. Floegel, A. *et al.* Identification of serum metabolites associated with risk of type 2 diabetes using a targeted metabolomic approach. *Diabetes* **62**, 639–648 (2013).
369. Toledo, E. *et al.* Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes Care* **39**, 833–846 (2016).
370. Wang, T. J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **17**, 448–453 (2011).
371. Xu, F. *et al.* Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. *J. Clin. Endocrinol. Metab.* **98**, E1060-5 (2013).
372. Lotta, L. A. *et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLoS Med.* **13**, 1–22 (2016).

373. Eastwood, S. V. *et al.* Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS One* **11**, (2016).
374. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
375. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
376. Beulens, J. W. J. *et al.* Alcohol consumption and risk of type 2 diabetes in European men and women: influence of beverage type and body size The EPIC-InterAct study. *J. Intern. Med.* **272**, 358–370 (2012).
377. Sacerdote, C. *et al.* Lower educational level is a predictor of incident type 2 diabetes in European countries: the EPIC-InterAct study. *Int. J. Epidemiol.* **41**, 1162–1173 (2012).
378. The InterAct Consortium. Long-Term Risk of Incident Type 2 Diabetes and Measures of Overall and Regional Obesity: The EPIC-InterAct Case-Cohort Study. *PLOS Med.* **9**, e1001230 (2012).
379. The InterAct Consortium. Mediterranean Diet and Type 2 Diabetes Risk in the European Prospective Investigation Into Cancer and Nutrition (EPIC) Study. *Diabetes Care* **34**, 1913 LP – 1918 (2011).
380. Ekelund, U. *et al.* Physical activity reduces the risk of incident type 2 diabetes in general and in abdominally lean and obese men and women: the EPIC-InterAct Study. *Diabetologia* **55**, 1944–1952 (2012).
381. Spijkerman, A. M. W. *et al.* Smoking and long-term risk of type 2 diabetes: The EPIC-InterAct study in European populations. *Diabetes Care* **37**, 3164–3171 (2014).
382. Mori, H. *et al.* Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8242–8247 (2002).
383. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
384. Taub, M. A. *et al.* Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed. *bioRxiv* 749010 (2019). doi:10.1101/749010
385. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
386. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, (2018).
387. Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat. Genet.* **51**, 51–62 (2019).
388. Langenberg, C. & Lotta, L. A. Genomic insights into the causes of type 2 diabetes. *Lancet* **391**, 2463–2474 (2018).
389. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
390. Ding, Z., Mangino, M., Aviv, A., Spector, T. & Durbin, R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* **42**, 7–10 (2014).
391. De Meyer, T. *et al.* Telomere Length as Cardiovascular Aging Biomarker. *J. Am. Coll. Cardiol.* **72**, 805–813 (2018).

392. Minamino, T. *et al.* Endothelial cell senescence in human atherosclerosis: role of telomere in endothelial dysfunction. *Circulation* **105**, 1541–4 (2002).
393. Daniali, L. *et al.* Telomeres shorten at equivalent rates in somatic tissues of adults. *Nat. Commun.* **4**, 1597 (2013).
394. Willeit, P. *et al.* Cellular aging reflected by leukocyte telomere length predicts advanced atherosclerosis and cardiovascular disease risk. *Arterioscler. Thromb. Vasc. Biol.* **30**, 1649–56 (2010).
395. De Meyer, T. *et al.* Systemic telomere length and preclinical atherosclerosis: the Asklepios Study. *Eur. Heart J.* **30**, 3074–3081 (2009).
396. Fernández-Alvira, J. M. *et al.* Short Telomere Load, Telomere Length, and Subclinical Atherosclerosis. *J. Am. Coll. Cardiol.* **67**, 2467–2476 (2016).
397. Bekaert, S. *et al.* Telomere length and cardiovascular risk factors in a middle-aged population free of overt cardiovascular disease. *Aging Cell* **6**, 639–647 (2007).
398. Benetos, A. *et al.* Tracking and fixed ranking of leukocyte telomere length across the adult life course. *Aging Cell* **12**, 615–621 (2013).
399. Park, J.-I. *et al.* Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature* **460**, 66–72 (2009).
400. Endorf, E. B. *et al.* Telomerase Reverse Transcriptase Deficiency Prevents Neointima Formation Through Chromatin Silencing of E2F1 Target Genes. *Arterioscler. Thromb. Vasc. Biol.* **37**, 301–311 (2017).
401. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science (80-.).* **347**, 81 LP – 83 (2015).
402. Wiktor, A. *et al.* Clinical significance of Y chromosome loss in hematologic disease. *Genes, Chromosom. Cancer* **27**, 11–16 (2000).
403. Crowe, F. L. *et al.* Fruit and vegetable intake and mortality from ischaemic heart disease: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Heart study. *Eur. Heart J.* **32**, 1235–1243 (2011).
404. Verhoeven, J. E. *et al.* Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. *Mol. Psychiatry* **19**, 895–901 (2014).
405. Zhao, S. & Fernald, R. D. Comprehensive Algorithm for Quantitative Real-Time Polymerase Chain Reaction. *J. Comput. Biol.* **12**, 1047–1064 (2005).
406. Ma, Q. *et al.* MAGI3 negatively regulates Wnt/beta-catenin signaling and suppresses malignant phenotypes of glioma cells. *Oncotarget* **6**, 35851–65 (2015).
407. Ma, Q. *et al.* MAGI3 Suppresses Glioma Cell Proliferation via Upregulation of PTEN Expression. *Biomed. Environ. Sci.* **28**, 502–9 (2015).
408. Dell'Angelica, E. C., Mullins, C. & Bonifacino, J. S. AP-4, a novel protein complex related to clathrin adaptors. *J. Biol. Chem.* **274**, 7278–7285 (1999).
409. Hirst, J., Bright, N. A., Rous, B. & Robinson, M. S. Characterization of a fourth adaptor-related protein complex. *Mol. Biol. Cell* **10**, 2787–2802 (1999).
410. Bauer, P. *et al.* Mutation in the AP4B1 gene cause hereditary spastic paraplegia type 47 (SPG47). *Neurogenetics* **13**, 73–76 (2012).
411. Barber, E. K., Dasgupta, J. D., Schlossman, S. F., Trevillyan, J. M. & Rudd, C. E. The CD4 and CD8 antigens are coupled to a protein-tyrosine kinase (p56lck) that phosphorylates the CD3 complex. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 3277–3281 (1989).
412. Iwashima, M., Irving, B. A., van Oers, N. S., Chan, A. C. & Weiss, A. Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science* **263**, 1136–1139 (1994).

413. Sturm, R. A., Cassady, J. L., Das, G., Romo, A. & Evans, G. A. Chromosomal structure and expression of the human OTF1 locus encoding the Oct-1 protein. *Genomics* **16**, 333–341 (1993).
414. Segil, N., Roberts, S. B. & Heintz, N. Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity. *Science* **254**, 1814–1816 (1991).
415. Roberts, S. B., Segil, N. & Heintz, N. Differential phosphorylation of the transcription factor Oct1 during the cell cycle. *Science* **253**, 1022–1026 (1991).
416. Schild-Poulter, C., Shih, A., Yarymowich, N. C. & Hache, R. J. G. Down-regulation of histone H2B by DNA-dependent protein kinase in response to DNA damage through modulation of octamer transcription factor 1. *Cancer Res.* **63**, 7197–7205 (2003).
417. Wysocka, J. & Herr, W. The herpes simplex virus VP16-induced complex: the makings of a regulatory switch. *Trends Biochem. Sci.* **28**, 294–304 (2003).
418. Lupo, B. & Trusolino, L. Inhibition of poly(ADP-ribosyl)ation in cancer: Old and new paradigms revisited. *Biochim. Biophys. Acta - Rev. Cancer* **1846**, 201–215 (2014).
419. Déjardin, J. & Kingston, R. E. Purification of Proteins Associated with Specific Genomic Loci. *Cell* **136**, 175–186 (2009).
420. Liang, Y. *et al.* Association of ACYP2 and TSPYL6 Genetic Polymorphisms with Risk of Ischemic Stroke in Han Chinese Population. *Mol. Neurobiol.* **54**, 5988–5995 (2017).
421. Liu, M. *et al.* Association between single nucleotide polymorphisms in the TSPYL6 gene and breast cancer susceptibility in the Han Chinese population. *Oncotarget* **7**, 54771–54781 (2016).
422. Boulay, J. L., Dennefeld, C. & Alberga, A. The Drosophila developmental gene snail encodes a protein with nucleic acid binding fingers. *Nature* **330**, 395–398 (1987).
423. Hay, R. T. SUMO: A History of Modification. *Mol. Cell* **18**, 1–12 (2005).
424. Jones, a. M. *et al.* TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. *Gut* **61**, 248–254 (2012).
425. Lührig, S. *et al.* Lrrc34, a novel nucleolar protein, interacts with npm1 and ncl and has an impact on pluripotent stem cells. *Stem Cells Dev.* **23**, 2862–74 (2014).
426. Fingerlin, T. E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat. Genet.* **45**, 613–620 (2013).
427. Chow, A., Hao, Y. & Yang, X. Molecular characterization of human homologs of yeast MOB1. *Int. J. cancer* **126**, 2079–2089 (2010).
428. Lai, Z.-C. *et al.* Control of cell proliferation and apoptosis by mob as tumor suppressor, mats. *Cell* **120**, 675–685 (2005).
429. Kerjan, G. *et al.* Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 6766–6771 (2009).
430. Kiss, T., Fayet-Lebaron, E. & Jády, B. E. Box H/ACA Small Ribonucleoproteins. *Mol. Cell* **37**, 597–606 (2010).
431. Kwak, J. E., Wang, L., Ballantyne, S., Kimble, J. & Wickens, M. Mammalian GLD-2 homologs are poly(A) polymerases. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4407–4412 (2004).
432. Glahder, J. A. & Norrild, B. Involvement of hGLD-2 in cytoplasmic polyadenylation of human p53 mRNA. *APMIS* **119**, 769–775 (2011).
433. Wyman, S. K. *et al.* Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome*

Res. **21**, 1450–1461 (2011).
434. Schmidt, C. K. *et al.* Systematic E2 screening reveals a UBE2D-RNF138-CtIP axis promoting DNA repair. *Nat. Cell Biol.* **17**, 1458–1470 (2015).
435. Lehner, B. *et al.* Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics* **83**, 153–167 (2004).
436. Tang, W., Kannan, R., Blanchette, M. & Baumann, P. Telomerase RNA biogenesis involves sequential binding by Sm and Lsm complexes. *Nature* **484**, 260–264 (2012).
437. Baumann, P. Pot1, the Putative Telomere End-Binding Protein in Fission Yeast and Humans. *Science (80-.).* **292**, 1171–1175 (2001).
438. Hockemeyer, D. & Collins, K. Control of telomerase action at human telomeres. *Nat. Struct. Mol. Biol.* **22**, 848–852 (2015).
439. Lange, T. De. Shelterin : the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100–2110 (2005).
440. Shimizu, A. *et al.* A novel giant gene CSMD3 encoding a protein with CUB and sushi multiple domains: a candidate gene for benign adult familial myoclonic epilepsy on human chromosome 8q23.3-q24.1. *Biochem. Biophys. Res. Commun.* **309**, 143–154 (2003).
441. Toomes, C. *et al.* The presence of multiple regions of homozygous deletion at the CSMD1 locus in oral squamous cell carcinoma question the role of CSMD1 in head and neck carcinogenesis. *Genes. Chromosomes Cancer* **37**, 132–140 (2003).
442. Scholnick, S. B. & Richter, T. M. The role of CSMD1 in head and neck carcinogenesis. *Genes, chromosomes & cancer* **38**, 281–283 (2003).
443. Otsuka, M., Mizuno, Y., Yoshida, M., Kagawa, Y. & Ohta, S. Nucleotide sequence of cDNA encoding human cytochrome c oxidase subunit VIc. *Nucleic Acids Res.* **16**, 10916 (1988).
444. Kile, B. T. *et al.* The SOCS box: a tale of destruction and degradation. *Trends Biochem. Sci.* **27**, 235–241 (2002).
445. Chen, L.-Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).
446. Chang, C.-W., Hsu, W.-B., Tsai, J.-J., Tang, C.-J. C. & Tang, T. K. CEP295 interacts with microtubules and is required for centriole elongation. *J. Cell Sci.* **129**, 2501–2513 (2016).
447. Wu, X. *et al.* ATM phosphorylation of Nijmegen breakage syndrome protein is required in a DNA damage response. *Nature* **405**, 477 (2000).
448. Banin, S. *et al.* Enhanced Phosphorylation of p53 by ATM in Response to DNA Damage. *Science (80-.).* **281**, 1674 LP – 1677 (1998).
449. Fan, J. *et al.* Tetrameric Acetyl-CoA Acetyltransferase 1 Is Important for Tumor Growth. *Mol. Cell* **64**, 859–874 (2016).
450. Fukao, T. *et al.* Molecular cloning and sequence of the complementary DNA encoding human mitochondrial acetoacetyl-coenzyme A thiolase and study of the variant enzymes in cultured fibroblasts from patients with 3-ketothiolase deficiency. *J. Clin. Invest.* **86**, 2086–2092 (1990).
451. Liu, L. *et al.* MCAF1/AM is involved in Sp1-mediated maintenance of cancer-associated telomerase activity. *J. Biol. Chem.* **284**, 5165–5174 (2009).
452. Liu, L. *et al.* MCAF1/AM Is Involved in Sp1-mediated Maintenance of Cancer-associated Telomerase Activity. *J. Biol. Chem.* **284**, 5165–5174 (2009).

453. Lee, J. & Zhou, P. DCAFs, the Missing Link of the CUL4-DDB1 Ubiquitin Ligase. *Mol. Cell* **26**, 775–780 (2007).
454. Gao, J. *et al.* The CUL4-DDB1 ubiquitin ligase complex controls adult and embryonic stem cell differentiation and homeostasis. *Elife* **4**, (2015).
455. Axe, E. L. *et al.* Autophagosome formation from membrane compartments enriched in phosphatidylinositol 3-phosphate and dynamically connected to the endoplasmic reticulum. *J. Cell Biol.* **182**, 685 LP – 701 (2008).
456. Shen, Z., Huang, S., Fang, M. & Wang, X. ENTPD5, an Endoplasmic Reticulum UDPase, Alleviates ER Stress Induced by Protein Overloading in AKT-Activated Cancer Cells. *Cold Spring Harb. Symp. Quant. Biol.* **76**, 217–223 (2011).
457. Fang, M. *et al.* The ER UDPase ENTPD5 Promotes Protein N-Glycosylation, the Warburg Effect, and Proliferation in the PTEN Pathway. *Cell* **143**, 711–724 (2010).
458. Heeringa, S. F. *et al.* COQ6 mutations in human patients produce nephrotic syndrome with sensorineural deafness. *J. Clin. Invest.* **121**, 2013–2024 (2011).
459. Tsang, W. Y. *et al.* CP110 Cooperates with Two Calcium-binding Proteins to Regulate Cytokinesis and Genome Stability. *Mol. Biol. Cell* **17**, 3423–3434 (2006).
460. Hayashi, R., Goto, Y., Ikeda, R., Yokoyama, K. K. & Yoshida, K. CDCA4 is an E2F transcription factor family-induced nuclear factor that regulates E2F-dependent transcriptional activation and cell proliferation. *J. Biol. Chem.* **281**, 35633–35648 (2006).
461. Kranz, T. M. *et al.* The chromosome 15q14 locus for bipolar disorder and schizophrenia: is C15orf53 a major candidate gene? *J. Psychiatr. Res.* **46**, 1414–1420 (2012).
462. Ebinu, J. O. *et al.* RasGRP links T-cell receptor signaling to Ras. *Blood* **95**, 3199–3203 (2000).
463. Roose, J. P., Mollenauer, M., Gupta, V. A., Stone, J. & Weiss, A. A diacylglycerol-protein kinase C-RasGRP1 pathway directs Ras activation upon antigen receptor stimulation of T cells. *Mol. Cell. Biol.* **25**, 4426–4441 (2005).
464. van der Velden, L. M. *et al.* Heteromeric interactions required for abundance and subcellular localization of human CDC50 proteins and class 1 P4-ATPases. *J. Biol. Chem.* **285**, 40088–40096 (2010).
465. Paulusma, C. C. & Oude Elferink, R. P. J. The type 4 subfamily of P-type ATPases, putative aminophospholipid translocases with a role in human disease. *Biochim. Biophys. Acta* **1741**, 11–24 (2005).
466. Gao, L. *et al.* Identification of Rare Variants in ATP8B4 as a Risk Factor for Systemic Sclerosis by Whole-Exome Sequencing. *Arthritis Rheumatol.* **68**, 191–200 (2016).
467. Hosford, D. *et al.* Candidate Single-Nucleotide Polymorphisms From a Genomewide Association Study of Alzheimer Disease. *JAMA Neurol.* **65**, 45–53 (2008).
468. Palfreyman, M. T. & Jorgensen, E. M. Unc13 Aligns SNAREs and Superprimes Synaptic Vesicles. *Neuron* **95**, 473–475 (2017).
469. McRory, J. E. *et al.* Molecular and functional characterization of a family of rat brain T-type calcium channels. *J. Biol. Chem.* **276**, 3999–4011 (2001).
470. Cribbs, L. L. *et al.* Cloning and characterization of alpha1H from human heart, a member of the T-type Ca2+ channel gene family. *Circ. Res.* **83**, 103–109 (1998).
471. Daniil, G. *et al.* CACNA1H Mutations Are Associated With Different Forms of Primary Aldosteronism. *EBioMedicine* **13**, 225–236 (2016).
472. Vitko, I. *et al.* Functional Characterization and Neuronal Modeling of the Effects of

Childhood Absence Epilepsy Variants of CACNA1H, a T-Type Calcium Channel. *J. Neurosci.* **25**, 4844–4855 (2005).
473. Van Steensel, B., Smogorzewska, A. & De Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).
474. Tian, Y. *et al.* C. elegans Screen Identifies Autophagy Genes Specific to Multicellular Organisms. *Cell* **141**, 1042–1055 (2010).
475. Smogorzewska, A. *et al.* Control of human telomere length by TRF1 and TRF2. *Mol. Cell. Biol.* **20**, 1659–68 (2000).
476. Inano, S. *et al.* RFWD3-Mediated Ubiquitination Promotes Timely Removal of Both RPA and RAD51 from DNA Damage Sites to Facilitate Homologous Recombination. *Mol. Cell* **66**, 622-634.e8 (2017).
477. Fu, X. *et al.* RFWD3-Mdm2 ubiquitin ligase complex positively regulates p53 stability in response to DNA damage. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 4579–4584 (2010).
478. Lehner, B. & Sanderson, C. M. A protein interaction framework for human mRNA degradation. *Genome Res.* **14**, 1315–1323 (2004).
479. Shintani, M., Urano, M., Takakuwa, Y., Kuroda, M. & Kamoshida, S. Immunohistochemical characterization of pyrimidine synthetic enzymes, thymidine kinase-1 and thymidylate synthase, in various types of cancer. *Oncol. Rep.* **23**, 1345–1350 (2010).
480. Tempel, W. *et al.* Nicotinamide riboside kinase structures reveal new pathways to NAD+. *PLoS Biol.* **5**, e263 (2007).
481. Han, Z. G. *et al.* Molecular cloning of six novel Krüppel-like zinc finger genes from hematopoietic cells and identification of a novel transregulatory domain KRNB. *J. Biol. Chem.* **274**, 35741–8 (1999).
482. Kotenko, S. V *et al.* IFN-lambdas mediate antiviral protection through a distinct class II cytokine receptor complex. *Nat. Immunol.* **4**, 69–77 (2003).
483. Prosser, H. M. *et al.* Prokineticin receptor 2 (Prokr2) is essential for the regulation of circadian behavior by the suprachiasmatic nuclei. *Proc. Natl. Acad. Sci.* **104**, 648 LP – 653 (2007).
484. Dodé, C. & Rondard, P. PROK2/PROKR2 Signaling and Kallmann Syndrome. *Front. Endocrinol. (Lausanne).* **4**, 19 (2013).
485. Zhu, L. *et al.* Inhibition of cell proliferation by p107, a relative of the retinoblastoma protein. *Genes Dev.* **7**, 1111–1125 (1993).
486. Ryoo, J. *et al.* The ribonuclease activity of SAMHD1 is required for HIV-1 restriction. *Nat. Med.* **20**, 936–941 (2014).
487. Laguette, N. *et al.* SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature* **474**, 654–657 (2011).
488. Margalef, P. *et al.* Stabilization of Reversed Replication Forks by Telomerase Drives Telomere Catastrophe. *Cell* **172**, 439-453.e14 (2018).
489. Ballew, B. J. *et al.* A recessive founder mutation in regulator of telomere elongation helicase 1, RTEL1, underlies severe immunodeficiency and features of Hoyeraal Hreidarsson syndrome. *PLoS Genet.* **9**, e1003695 (2013).
490. Stuart, B. D. *et al.* Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat. Genet.* **47**, 512 (2015).
491. Zhang, Y. *et al.* Overexpression of SCLIP promotes growth and motility in glioblastoma cells. *Cancer Biol. Ther.* **16**, 97–105 (2015).
492. You, R. *et al.* Apoptosis of dendritic cells induced by decoy receptor 3 (DcR3). **111**,

1480–1489 (2019).
493. Pitti, R. M. *et al.* Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396**, 699–703 (1998).
494. Yang, C.-R. *et al.* Soluble decoy receptor 3 induces angiogenesis by neutralization of TL1A, a cytokine belonging to tumor necrosis factor superfamily and exhibiting angiostatic action. *Cancer Res.* **64**, 1122–1129 (2004).
495. Chevrier, S. & Corcoran, L. M. BTB-ZF transcription factors, a growing family of regulators of early and late B-cell development. *Immunol. Cell Biol.* **92**, 481–8 (2014).
496. Chen, W.-Y. *et al.* Inhibition of the androgen receptor induces a novel tumor promoter, ZBTB46, for prostate cancer metastasis. *Oncogene* **36**, 6213 (2017).
497. Li, J. S. Z. *et al.* TZAP: A telomere-associated protein involved in telomere length control. *Science (80-.).* **355**, 638–641 (2017).
498. Jahn, A. *et al.* ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep.* **18**, 929–946 (2017).
499. Adamson, B., Smogorzewska, A., Sigoillot, F. D., King, R. W. & Elledge, S. J. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.* **14**, 318–328 (2012).

Appendix B

List of publications authored and co-authored during PhD, including manuscript under review or in preparation.

Li C*, Stoma S*, Lotta LA*, Warner S*, Albrecht E, Allione A, Arp PP, Broer L, Buxton JL, Couto Alves A, Deelen J, Fedko IO, Gordon SD, Jiang T, Kerrison N, Karlsson R, Loe TK, Massimo M, Milaneschi Y, Miraglio B, Pervjakova N, Russo A, Surakka I, van der Spek A, Verhoeven JE, Amin N, Beekman M, Blakemore AI, Canzian F, Hamby SE, Hottenga JJ, Jones PD, Jousilahti P, Mägi R, Medland SE, Montgomery GW, Nyholt DR, Perola M, Saloma V, Suchiman HE, van Heemst D, Willemsen G, Agudo A, Boeing H, Boomsma DI, Chirlaque MD, Fagherazzi G, Ferrari P, Franks P, Gieger C , Eriksson JG, Gunter M, Hägg S, Hovatta I, Imaz L, Kaaks R, Key T, Krogh V, Martin NG, Melander O, Metspalu A, Moreno C, Onland-Moret CN, Nilsson P, Ong KK, Overvad K, Palli D, Panico S, Pedersen NL, Penninx BWJH, Quirós JR, Jarvelin MR, Rodríguez-Barranco M, Scott RA, Severi G, Slagboom EP, Spector TD, Tjonneland A, Trichopoulou A, Tumino R, Uitterlinden AG, van der Schouw Y, van Duijn CM, Weiderpass E, Denchi EL, Matullo G, Butterworth AS , Danesh J , Samani NJ, Wareham NJ, Nelson CP, Langenberg C*, Codd V*. "Genetic analysis links nucleotide metabolism to leukocyte telomere length". 2019 (*Manuscript under review*, *Presented at the American Society of Human Genetics Annual Meeting, 2018)*

Li C, Lotta LA, Zuber V, Stewart ID, Scott RA, Wareham NJ, Burgess S, Langenberg C. "Characterizing gene-specific associations of LDL cholesterol with type 2 diabetes using untargeted metabolomics". 2018 (*Manuscript in preparation*, *Presented at the International Mendelian Randomization Conference, Bristol, 2017)*

Podmore C, Stewart ID, …, **Li C**, …, Langenberg C. "Genetic regulation of iron metabolism, chronic iron overload and iron-iissue deposition in non-HFE carriers". 2019 (*Manuscript in preparation*)

Lotta LA*, Mokrosiński J*, Mendes de Oliveira E*, **Li C**, Sharp SJ, Luan J, Brouwers B, Ayinampudi V, Bowker N, D. Stewart ID, Wheeler E, Day FR, Perry JRB, Langenberg C*, Wareham NJ*, Farooqi IS*. "Human gain-of-function MC4R variants exhibit signalling bias and protect against obesity". *Cell* 177.3 (2019): 597-607.e9

Surendran P*, Stewart ID*, …, **Li C**, …, Butterworth AS*, Langenberg C*. "Genetic architecture of human 'chemical individuality'". 2019 (*Manuscript in preparation*, *Presented at the American Society of Human Genetics Annual Meeting, 2018*)

Lotta LA, …, **Li C**, …, Langenberg C. "Genomic interconnectivity and phenotypic landscape at 144 metabolite-associated loci". 2019 (*Manuscript in preparation*, *Presented at the American Society of Human Genetics Annual Meeting, 2018*)

Lotta LA, Wittemans LBL, Zuber V, Stewart ID, Sharp SJ, Luan J, Day FR, **Li C**, Bowker N, Cai L, Rolfe EDL, Khaw KT, Perry JRB, O'Rahilly S, Scott RA, Savage DB, Burgess S, Wareham NJ, Langenberg C. "Specific genetic determinants of gluteofemoral versus abdominal fat

distribution and risk of type 2 diabetes and coronary disease". *JAMA.* 320.24 (2018): 2553-2563.

Lotta LA, Stewart ID, Sharp SJ, Day FR, Burgess S, Luan J, Bowker N, Cai L, **Li C**, Wittemans LBL, Kerrison ND, Khaw KT, McCarthy MI, O'Rahilly S, Scott RA, Savage DB, Perry JRB, Langenberg C, Wareham NJ. "Association of genetically enhanced lipoprotein lipase-mediated lipolysis and low-density lipoprotein cholesterol-lowering alleles with risk of coronary disease and type 2 diabetes." *JAMA Cardiol.* 3.10 (2018): 957-966.

Lotta LA, Dong L, **Li C**, Patel S, Stewart ID, Lim K, Day FR, Wheeler E, Glastonbury CA, Streek MV, Sharp SJ, Luan J, Bowker N, Schweiger M, Wittemans LBL, Kerrison ND, Cai L, Lucarelli DME, Barroso I, McCarthy MI, Scott RA, Zechner R, Perry JRB, Saudek V, Small KS, O'Rahilly S, Wareham NJ, Savage DB, Langenberg C. "Genome-wide scan and fine-mapping of rare nonsynonymous associations implicates intracellular lipolysis in fat distribution." *Biorxiv*. 10.1101/372128 (2018). *Under review*.